Geometric Gait Clustering for Unobtrusive Analysis

Grant Ellison
Dept. of Computer Science
Loyola Marymount University
Los Angeles, United States
gelliso1@lion.lmu.edu

Milla Penelope Markovic
Dept. of Computer Science
Loyola Marymount University
Los Angeles, United States
mmarkov1@lion.lmu.edu

Delaram Yazdansepas
Dept. of Computer Science
Loyola Marymount University
Los Angeles, United States
delaram@lmu.edu

2023 IEEE 19th International Conference on Body Sensor Networks (BSN) | 979-8-3503-3841-6/23/$31.00 ©2023 IEEE | DOI: 10.1109/BSN58485.2023.10331324

***Abstract*—An important field within Human Activity Recognition is the evaluation of disease and patient recovery through the assessment of gait patterns. Clustering has been used as a data-mining technique to find the prior patterns in subjects' gait patterns. Previous studies have shown the discriminative power of gait clustering on biometrics, and the ability to detect abnormal gait patterns and gait pathology. Previous techniques have relied on expensive machinery and closed environments for gait pattern extraction and/or simplistic featured approaches to clustering. Geometric time series clustering has developed in other fields as a method for incorporating the information from an entire time series sequence and comparing sequences with temporal distortions. We present a method for geometric gait clustering using accelerometer data from wearable sensors. Our methods include an approach to gait cycle averaging and a two-way clustering method for assessing the similarity of biometrics within gait cycle clusters. Our results demonstrate that our methods have significant discriminative efficacy for biometrics and may be a useful analytical tool for gait pathology.**

***Index Terms*—gait analysis, clustering, DTW, KMeans, machine learning, wearable sensors, gait pathology**

I. Introduction

Human Activity Recognition (HAR) is the classification and analysis of human activity by machines. The ubiquity of sensors in devices like smartwatches and smartphones has revolutionized activity recognition by obviating users' need to wear supplementary devices [1]. Triaxial accelerometers, which are pervasive in these devices, present a favorable foundation for real-time analysis due to their information-rich nature and ability to generate a time series signal. Gait analysis, which is a subset of HAR, has been widely employed in the fields of patient recovery and disease assessments [2], [3], [4], [5]. It is well established that gait patterns are influenced by various physical characteristics [6], [7].

Clustering is a data-mining technique employed to reveal the underlying relationships within a dataset. In the context of gait analysis, various methods have been deployed to cluster human gait patterns. These methods have been successfully applied in the detection of gait pathology as well as disease detection and prevention efforts [8], [9]. Other methods can be characterized in two main ways; first, featured approaches with manual extraction of features from the gait of subjects, and second, using computer vision for the detection of gait patterns and gait events from image recognition. [10] Presents gait clustering from several different gait detection methods, and can identify distinct biometric characteristics between clusters based on gait patterns. To do further analysis on the impact of biometrics on gait patterns, we combine accelerometer-based activity data with more complete subject biometrics data. Moreover, in contrast to [10], which is focused solely on the $K = 5$ and $K = 10$ number of clusters, we extend our analysis by reporting results for a broader range of cluster numbers, specifically $K = 2, 3, 4, 5$.

In other domains, time series clustering techniques have been employed to classify instances into clusters using spatiotemporal similarity [11]. Geometric approaches to time series clustering differ from feature-based approaches by considering the entire time series information instead of extracting specific features. However, to the best of our knowledge, the application of geometric time series clustering has not been previously explored in the context of clustering human gait patterns.

This study presents a system designed to untangle the influence of biometrics on gait patterns using accelerometer data captured by wearable devices. Our approach involves extracting gait patterns through peak detection for gait cycle identification and employing barycenter cycle averaging for gait cycle alignment. Our analysis follows a two-step procedure; first, we cluster based on the geometric shapes of the gait pattern for every subject. Second, we cluster the biometrics of each subject. Finally, we compare the cluster assignments of both steps to identify similar assignments of subjects to the same cluster. We use a within-sample probability assessment to measure the significance of our results. Our findings demonstrate that assigning subjects to clusters based on shapelets shows a considerable improvement over a random assignment, which illustrates the influence of biometrics on gait cycle patterns.

Our methods show discriminative power that replicates other studies. We assess the most impactful biometrics, and we find significant inter-cluster distinctions for subjects with past anterior cruciate ligament (ACL) reconstruction surgery. These results indicate that our methods hold promise for other researchers in analyzing gait pathology and monitoring patient recovery, all while leveraging the unobtrusive nature of a hip-worn accelerometer.

II. Methods

A. Data Collection

35 subjects (20 female, 15 male) wore three wearable sensors (ActiGraph GT9X) to collect triaxial accelerometer data (100 Hz, ± 16 g) for a series of five activities. Five subjects had prior ACL injuries. Subjects wore a sensor on their non-dominant wrist, one around their waist attached to an elastic belt on their non-dominant hip, and one on their non-dominant ankle. Each subject walked on the sidewalk for 90 meters, climbed up and down three flights of stairs, walked on the treadmill at 2.5 mph for two minutes, and jogged on the treadmill at 5.5 mph for one minute. For our clustering analysis, we use a total of 11 numerical biometrics and the self-reporting of ACL reconstructive surgery. Table I summarizes the biometrics used and sample statistics. We focus on the hip-worn accelerometers for analysis.

TABLE I
Sample Statistics

Features	Avg.	Std. Dev.
Age	27.8	12.0
Height (cm)	170.9	10.5
Weight (kg)	66.2	13.5
Bdoy Mass Index (BMI)	22.5	3.2
Dominant Leg Length (cm)	97.8	7.7
Shoe Size (US Men's)	8.1	2.4
Dominant Femur Length (cm)	51.3	4.4
Torso Length (cm)	42.9	5.0
Wingspan (cm)	173.9	14.7
Shoulder Circumference (cm)	105.8	10.5
Waist Circumference (cm)	86.5	12.1

B. Time Series Activity Data

From the trivariate time series signal, created by our triaxial accelerometers, we extract the vector magnitude of acceleration across three axes, *x, y, z*.

$$v_t = \sqrt{x_t^2 + y_t^2 + z_t^2} \tag{1}$$

This improves the robustness of our data against the orientation of the sensor and potential errors from differences in the fitting of sensors on our subjects. Activity data, $A = \{v_1, v_2, ..., v_T\}$, is a time series representation of the vector magnitude of acceleration, over time, sampled at 100hz. Figure 1 visualizes the time series signal of vector magnitude.

C. Gait Extraction

A gait is comprised of sequential gait cycles separated by gait events, most notably the heel-strike, and toe-off [2]. Our data was collected in a controlled environment, and labeled by our researchers, therefore a simple, rule-based algorithm is adequate to extract gait cycles. We use a peak-finding algorithm, PeakUtils, to detect gait events. Figure 2 demonstrates the system working on our example data. As shown in Figure 3, we construct a library of candidate subsequences $C_a = \{c_1, c_2, ..., c_L\}$ where c_l is a gait cycle extracted from the peak-indexing algorithm, and a denotes the activity and L is the number of gait cycles.

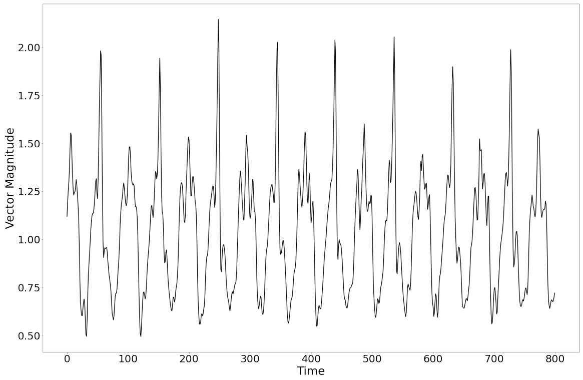

Fig. 1. Example of vector magnitude activity data.

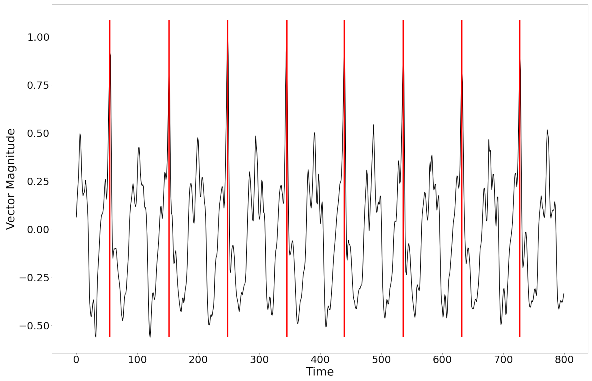

Fig. 2. The detection of gait events for walking sidewalk data.

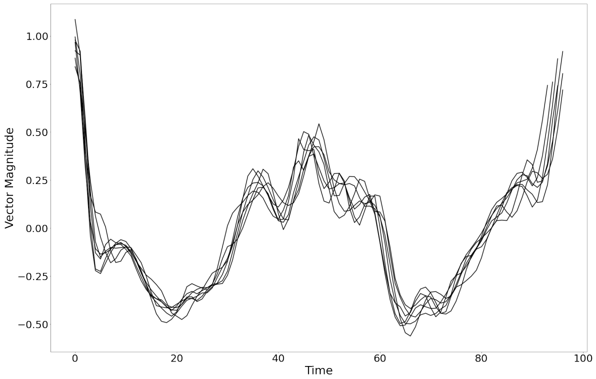

Fig. 3. All gait patterns from a single subject walking on the sidewalk.

D. Gait Averaging

Barycenter averaging is used to find the spatiotemporal sequence that minimizes the distance between a set of time series. We present a simple barycenter averaging technique we call gait averaging, shown in Figure 4, which follows two steps. Given our candidate library $C_a = \{c_1, c_2, ..., c_L\}$, the mean length, μ is found, and each candidate, c_l, is interpolated to a create a vector, $\vec{c_l}$, of length μ. Subsequently, the set of vectors is averaged to obtain a representative value.

$$gait_a = \frac{1}{L}\sum_{l=1}^{L}\vec{c_l} \tag{2}$$

where $\vec{c_l} \in \mathbb{R}^n$. Gait averaging is used again during the clustering process.

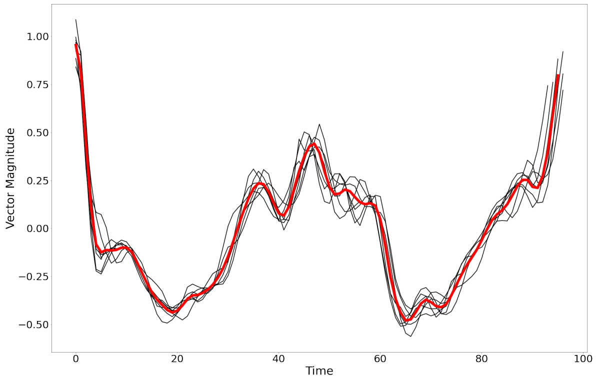

Fig. 4. Gait cycle averaging from the candidate library.

E. Gait Clustering

The goal of a clustering algorithm is to classify unlabelled data instances into clusters according to their similarity. We use Dynamic Time Warping (DTW) as a similarity metric for comparing time series data [12]. Its utility lies in addressing misalignment problems encountered when dealing with sequences that are out-of-phase or have varying durations. In the context of gait analysis, DTW has been extensively employed to quantify the spatiotemporal distance between gait cycles [2] [13]. We use clustering to identify the similarities present in our subjects' gait patterns. To cluster gait patterns, set $S = \{s_1, s_2, ..., s_N\}$ of gait patterns for each $subject_n$, is randomly sampled to create a set $M = \{m_1, m_2, ..., m_K\}$ of initial centroids where K is the number of clusters. Each $s_n \in S$ is compared by DTW to each $m_k \in M$, and assigned to the cluster C_k that results in the lowest DTW score. The clustering algorithm minimizes the inertia - the distance between each instance s_n and the centroid m_k of its assigned cluster C_k.

$$inertia = \sum_{n=1}^{N}\sum_{k=1}^{K} I(s_n \in C_k)DTW(s_n, m_k) \tag{3}$$

Where N is the number of gait patterns in S, K is the number of clusters, and I takes on the value of 1 if an gait pattern s_n is assigned to a cluster C_k, and a value of 0 otherwise. We use the *KMeans* algorithm, implemented with DTW and gait averaging, to iteratively update centroids, and assign membership of each s_n to a cluster C_k, according to Lloyd's algorithm [14]. To avoid local minima, we resample the initial centroids 10 times and return the cluster assignment and centroid set that resulted in the minimum inertia.

F. Biometrics Clustering

A set $B = \{\vec{b_1}, \vec{b_2}, ..., \vec{b_N}\}$ is created where each $\vec{b_n}$ is a vector containing a set of biometrics pertaining to $subject_n$. Centroids $M = \{\vec{m_1}, \vec{m_2}, ..., \vec{m_K}\}$, are initialized by randomly sampling the set B, K times. B is clustered with inertia defined as the sum of the Euclidean distances between each $\vec{b_n}$ and the centroid $\vec{m_k}$ of its assigned cluster C_k.

$$inertia = \sum_{n=1}^{N}\sum_{k=1}^{K} I(\vec{b_n} \in C_k)\sqrt{|\vec{b_n} - \vec{m_k}|^2} \tag{4}$$

where $\vec{c_l}, \vec{m_k} \in \mathbb{R}^n$, N is the number of subjects, K is the number of clusters, and I takes on the value of 1 if $\vec{b_n}$ is a member of cluster C_k, and a value of 0 otherwise. We again use the *KMeans* algorithm, now implemented with Euclidean distance and simple averaging - represented by equation 2 - to update centroids, and assign membership of each $\vec{b_n}$ to a cluster C_k. We resample the initial centroids 5 times and return the cluster assignment and centroid set that results in the minimum inertia.

III. Analysis and Results

A. Two-Way Clustering Similarity

To address the question of biometric impacts on the gait cycle pattern, we assess the similarity between the assignment of a $subject_n$ into the same cluster for their gait pattern and their biometric data. We use the common Adjusted Rand Index (ARI), to assess the improvement over the random assignment of clusters.

To assess the probability of achieving a particular result, we derive the $p - value$ in the following way; we consider each alternative biometric set within the K number of clusters, and measure the probability of achieving a result at least as high as observed. This method is an essential step in our analysis, but it comes with drawbacks. First, the assessment of the $p - value$ in this context measures the probability of obtaining a result higher than the next lowest value, rather than directly addressing the probability of achieving the specific value of interest. This distinction should be taken into account when interpreting the significance of the $p - value$. Second, it is important to recognize that small sample sizes pose challenges in interpreting results in a traditional sense. When the sample size is limited, the $p - value$ may not accurately reflect the true probability. While this latter drawback does not impact the validity of our current results, it could potentially affect datasets with fewer biometrics available for analysis.

B. Results

Table II is an account of our results for the *Walking Sidewalk - Natural Pace*. Here we show the ARI of several select biometric sets. Table II highlights the significant enhancements our methods offer compared to random similarity across various cluster numbers of K. It is worth noting that differences in the similarity between different K values are expected due

TABLE II
ARI AND $p - value$ OF SELECT BIOMETRICS

	K Clusters			
Features	K = 2	K = 3	K = 4	K = 5
ACL	***0.18	**0.04	0.00	0.01
Age	**0.14	*0.03	-0.03	0.01
Torso Length	*0.04	0.01	**0.06	***0.11
Age, Sex, BMI	**0.14	*0.03	0.00	**0.10
Age, Leg Length, Torso Length	***0.18	**0.03	-0.01	**0.09

* = $p - value$ below 10%
** = $p - value$ below 5%
*** = $p - value$ below 1%

to the variation in the number of natural clusters that might be present depending on the biometrics being evaluated.

C. Biometric Distributions

Table III shows the biometrics distributions for each of the three clusters. Our results are consistent with the findings of [10], where they show that the distribution of sex and body area both have significant inter-cluster distinctions. Age, sex, ACL reconstructive surgery, and height all provide notable distinctions between clusters.

TABLE III
$K = 3$ COMPARISON OF BIOMETRIC DISTRIBUTIONS

	Cluster		
Features	1	2	3
Sex (male)	33%	33%	45%
Age	32.7 ±17.2	21.6 ±4.7	26.6 ±9.7
Height (cm	173.2 ±10.0	162.0 ±7.0	171.2 ±10.8
Dom. Leg Length (cm)	100.5 ±5.9	94.3 ±5.1	97.1 ±8.4
Weight (kg)	63.0 ±11.2	56.3 ±4.4	68.7 ±14.5
BMI	20.9 ±2.2	21.4 ±0.7	23.3 ±3.4
Torso Length (cm)	42.2 ±4.3	39.5 ±3.7	43.7 ±5.4
Shoe Size (US Mens')	8.4 ±2.6	6.7 ±3.3	8.2 ±2.2
ACL	0%	0%	22%

avg. $\pm std.dev.$

D. Gait Pathology

Among our self-reported biometrics, we consider previous ACL reconstructive surgery. Out of the 35 subjects in our study, 5 reported having undergone ACL reconstructive surgery. By employing our gait clustering algorithm and considering two and three clusters ($K = 2$ and $K = 3$), we observe that all subjects who had previous ACL surgery were assigned to the same cluster, specifically for the *Walking Sidewalk - Natural Pace* activity. Notably, this outcome is not evident in the case of the *Walking Treadmill - 2.5 MPH* activity, possibly attributable to the prescribed pace enforced by the treadmill.

IV. DISCUSSION

Our findings provide clear evidence supporting the intuitive notion that gait cycle patterns are influenced, to some degree, by an individual's physical attributes. The noticeable improvement over the random assignment, between geometric gait pattern clusters and biometric clusters, highlights the significant role played by these physical characteristics in shaping gait patterns. Specifically, our results utilizing ACL reconstruction data underscore the value of gait clustering through accelerometers in facilitating unobtrusive analysis of gait abnormalities and monitoring patient recovery. The sample size of 35 subjects is one area that limits the generalizability of our work. While this sample size may adequately demonstrate the efficacy of our methods, a small sample size makes it difficult to reach conclusive results regarding the impact of biometrics on gait, and gait abnormality. Further application of our methods to additional datasets presents an opportunity to assess the discriminative capabilities of our methods, for diseases and disorders including Parkinson's disease and cerebral palsy, and for post-operative recovery. This exploration holds promise in evaluating the effectiveness of our techniques in these contexts and expanding the scope of their potential applications.

REFERENCES

[1] O. D. Lara and M. A. Labrador, "A survey on human activity recognition using wearable sensors," vol. 15, pp. 1192–1209. Conference Name: IEEE Communications Surveys & Tutorials.

[2] S. R. Dandu, M. M. Engelhard, M. D. Goldman, and J. Lach, "Determining physiological significance of inertial gait features in multiple sclerosis," in *2016 IEEE 13th International Conference on Wearable and Implantable Body Sensor Networks (BSN)*, pp. 266–271, IEEE.

[3] R. Baker, "Gait analysis methods in rehabilitation," vol. 3, no. 1, p. 4.

[4] R. D. Gurchiek, R. H. Choquette, B. D. Beynnon, J. R. Slauterbeck, T. W. Tourville, M. J. Toth, and R. S. McGinnis, "Remote gait analysis using wearable sensors detects asymmetric gait patterns in patients recovering from ACL reconstruction," in *2019 IEEE 16th International Conference on Wearable and Implantable Body Sensor Networks (BSN)*, pp. 1–4, IEEE.

[5] G. Cicirelli, D. Impedovo, V. Dentamaro, R. Marani, G. Pirlo, and T. R. D'Orazio, "Human gait analysis in neurodegenerative diseases: A review," vol. 26, no. 1, pp. 229–242. Conference Name: IEEE Journal of Biomedical and Health Informatics.

[6] M. O. Derawi, "Accelerometer-based gait analysis, a survey,"

[7] S. Aghabozorgi, A. Seyed Shirkhorshidi, and T. Ying Wah, "Time-series clustering – a decade review," vol. 53, pp. 16–38.

[8] A. Nguyen, N. Roth, N. H. Ghassemi, J. Hannink, T. Seel, J. Klucken, H. Gassner, and B. M. Eskofier, "Development and clinical validation of inertial sensor-based gait-clustering methods in parkinson's disease," vol. 16, no. 1, p. 77.

[9] E. Dolatabadi, A. Mansfield, K. K. Patterson, B. Taati, and A. Mihailidis, "Mixture-model clustering of pathological gait patterns," vol. 21, no. 5, pp. 1297–1305. Conference Name: IEEE Journal of Biomedical and Health Informatics.

[10] B. DeCann, A. Ross, and M. Culp, "On clustering human gait patterns," in *2014 22nd International Conference on Pattern Recognition*, pp. 1794–1799. ISSN: 1051-4651.

[11] "Clustering time series using unsupervised-shapelets | IEEE conference publication | IEEE xplore."

[12] P. Senin, "Dynamic time warping algorithm review,"

[13] M. Engelhard, S. R. Dandu, J. Lach, M. Goldman, and S. Patek, "Toward detection and monitoring of gait pathology using inertial sensors under rotation, scale, and offset invariant dynamic time warping," in *Proceedings of the 10th EAI International Conference on Body Area Networks*, ICST.

[14] S. Lloyd, "Least squares quantization in PCM," vol. 28, no. 2, pp. 129–137.