An Interpretable Distance Measure for Multivariate Non-Stationary Physiological Signals

Sylvain W. Combettes, Charles Truong, and Laurent Oudre

Université Paris-Saclay, Université Paris Cité, ENS Paris-Saclay, CNRS, SSA, INSERM, Centre Borelli

Introduction

Motivation: study of human locomotion [3]

- Angular velocity recorded on the left and right feet using a pair of sensors.
- Protocol: standing, walking, turning around, walking back, and standing.
- Multivariate signals with $d = 16$ dimensions: norms of the STFT (Short Time Fourier Transform) of each univariate foot recording.

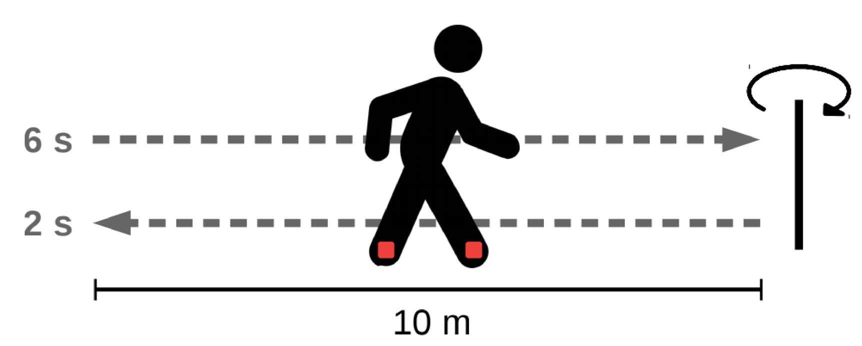

Figure 1: Gait signal acquisition protocol.

Comparing multivariate time series

- Popular distances (Euclidean distance, Dynamic Time Warping) can not handle non-stationarity.
- Our distance is interpretable and can compare non-stationary signals: (i) symbolization, (ii) distance on strings.

Symbolization technique

1. Segmentation step: a real-valued signal of length n is split into w segments ($w < n$).
2. Quantization step: each segment is mapped to a discrete value taken from a set of A symbols.

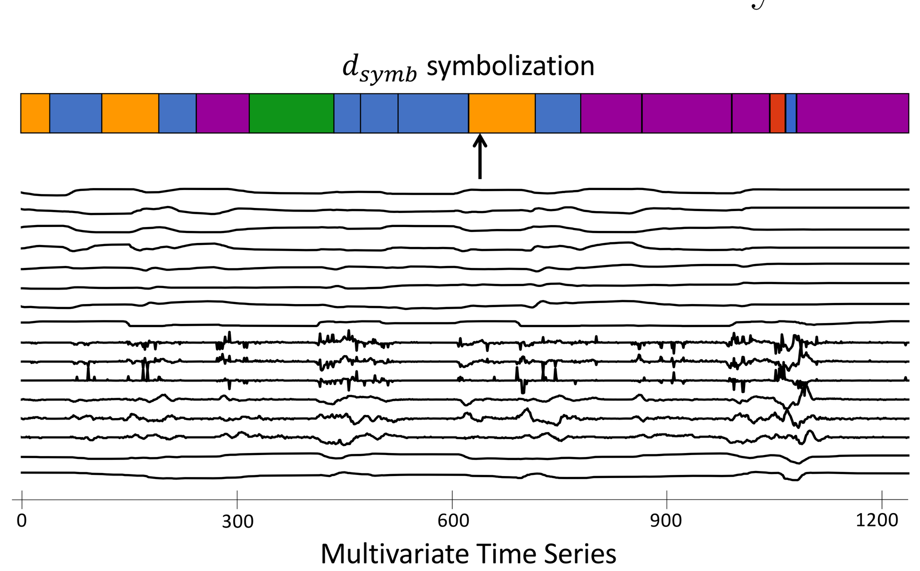

Figure 2: Symbolization of a multivariate signal into a string called a symbolic sequence (visualized using a color bar).

Method: the d_symb symbolization and distance measure

Steps of d_{symb} [1]

1. Segmentation: change-point detection (on the mean).
2. Quantization: K-means clustering (of the means per segment), with $K = A$.
3. Distance: general edit distance between the resulting symbolic signals.

Change-point detection

Finding the w^* unknown instants $t_1^* < t_2^* < \ldots < t_{w^*+1}^*$ where the mean of signal $x = (x_1, \ldots, x_n)$ change abruptly:

$$(\hat{w}, \hat{t}_1, \ldots, \hat{t}_{\hat{w}+1}) = \underset{(w, t_1, \ldots, t_{w+1})}{\arg\min} \sum_{k=0}^{w+1} \sum_{t=t_k}^{t_{k+1}-1} \|x_t - \bar{x}_{t_k:t_{k+1}}\|^2 + \lambda w,$$

where $\bar{x}_{t_k:t_{k+1}}$ is the empirical mean of $\{x_{t_k}, \ldots, x_{t_{k+1}-1}\}$ and $\lambda > 0$ is a penalization parameter.

- Compromise between the reconstruction error and the number of change-points.
- When λ is small, many change-points are detected. For calibration purposes, we use $\lambda = \ln(n)$.
- Solved using the Pruned Exact Linear Time (PELT) algorithm [2], which is shown to have $\mathcal{O}(n)$ complexity (under some assumptions).

Levering the general edit distance

1. Preprocessing.
   - Including the segment length information: replicating each symbol proportionally to its segment length. Example: `abd` becomes `aabbbbdd`.
   - Shortening: dividing each length by the minimum length. Example: `aabbbbdd` becomes `abbd`.
2. Applying the general edit distance with custom costs.
   - Edit distance on strings (a.k.a Levenshtein distance): minimal cost of a sequence of operations that transform a string into another.
   - Allowed simple operations and their costs:
     - Substitution: Euclidean distance between the cluster centers of the symbols.
     - Insertion: max of substitution costs.
     - Deletion: max of substitution costs.
   - Total cost: sum of the costs of the simple operations.

Experimental results: interpretation of the symbolization

The general structure is coherent with the clinical protocol: change-point detection finds stationary segments, and each symbol can be associated with a specific type of behavior.

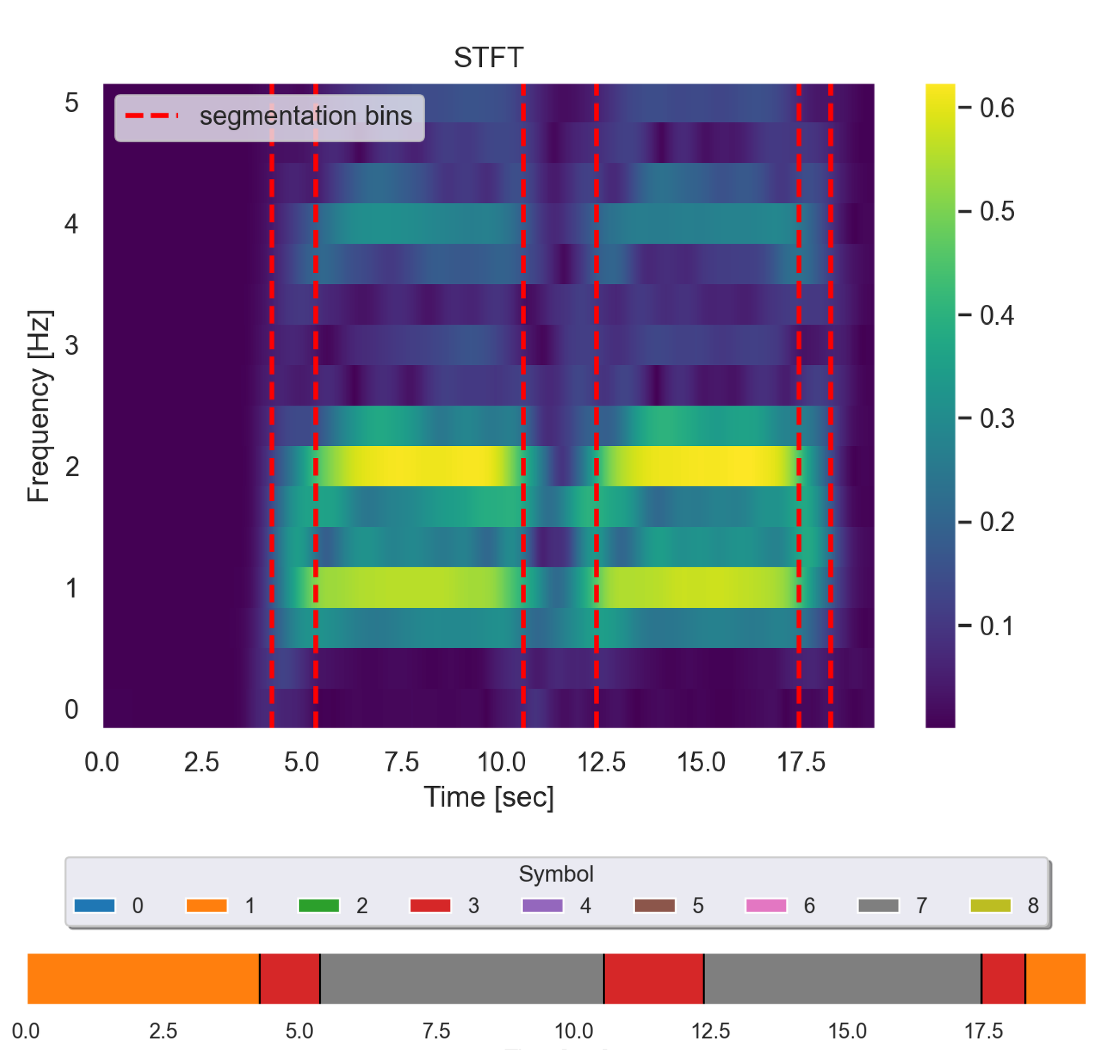

Figure 3: Spectrogram (multivariate signal) of one recording with its corresponding color bar.

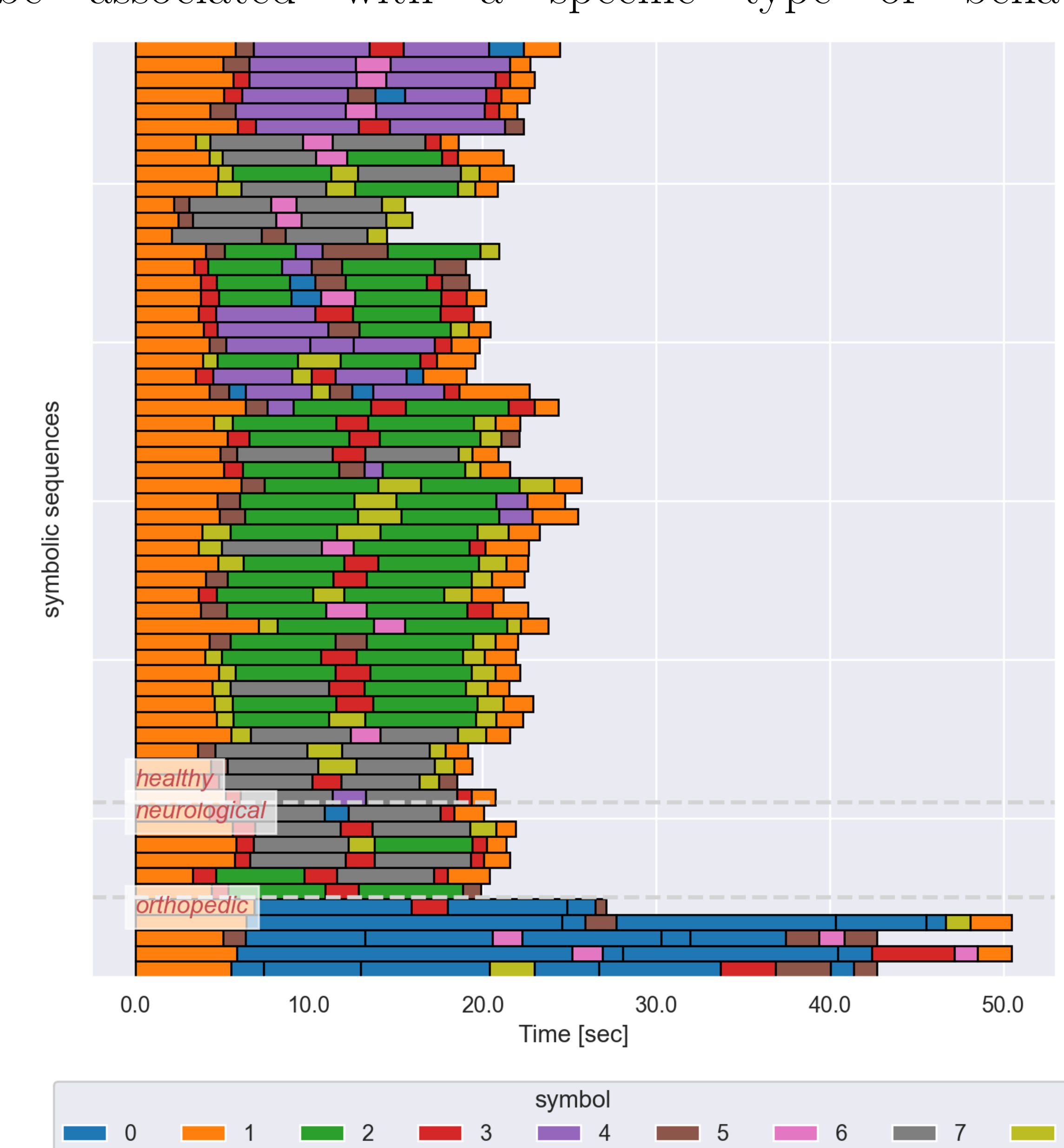

Figure 4: Color bars for 60 recordings, with 3 classes and $A = 9$.

Experimental results: the distance

Silhouette coefficient calculated using the distance matrix and the ground truth patient group:

Distance measure	Mean Silhouette score	Median Silhouette score
DTW-D	0.15	0.18
DTW-I	0.15	0.19
d_{symb}	0.33	0.40

Robustness to the difference in lengths:

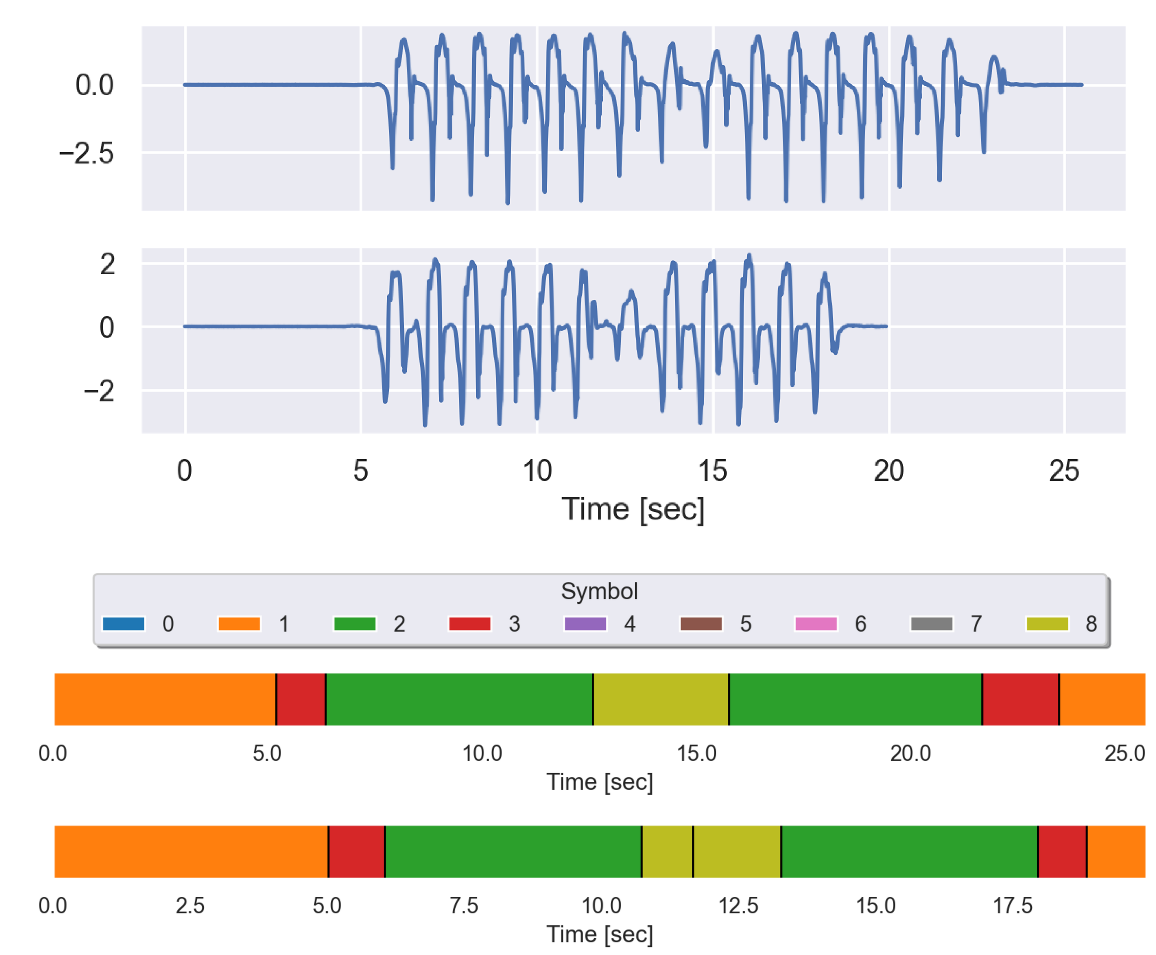

Figure 5: Two gait signals and their color bars.

References

[1] S. W. Combettes, C. Truong, and L. Oudre. An interpretable distance measure for multivariate non-stationary physiological signals. In *2023 IEEE International Conference on Data Mining Workshops (ICDMW)*, pages 533–539, 2023.

[2] R. Killick, P. Fearnhead, and I. A. Eckley. Optimal detection of changepoints with a linear computational cost. *Journal of the American Statistical Association*, 107(500):1590–1598, 2012.

[3] C. Truong and et al. A Data Set for the Study of Human Locomotion with Inertial Measurement Units. *Image Processing On Line*, 9:381–390, 2019. https://doi.org/10.5201/ipol.2019.265.

sylvain.combettes@ens-paris-saclay.fr
https://sylvaincom.github.io