Discovering Optimal Variable-length Time Series Motifs in Large-Scale Wearable Recordings of Human Bio-behavioral Signals

Tiantian Feng[1], Shrikanth Narayanan[1]

[1]Signal Analysis and Interpretation Lab (SAIL), University of Southern California

USC University of Southern California

SAIL

Summary

Background:

- Motifs are repetitive similar patterns that frequently appear in time-series.
- Motifs existing in wearable sensor signals can help to understand bio-behavioral patterns such as sleep pattern, commute behavior.

Challenges:

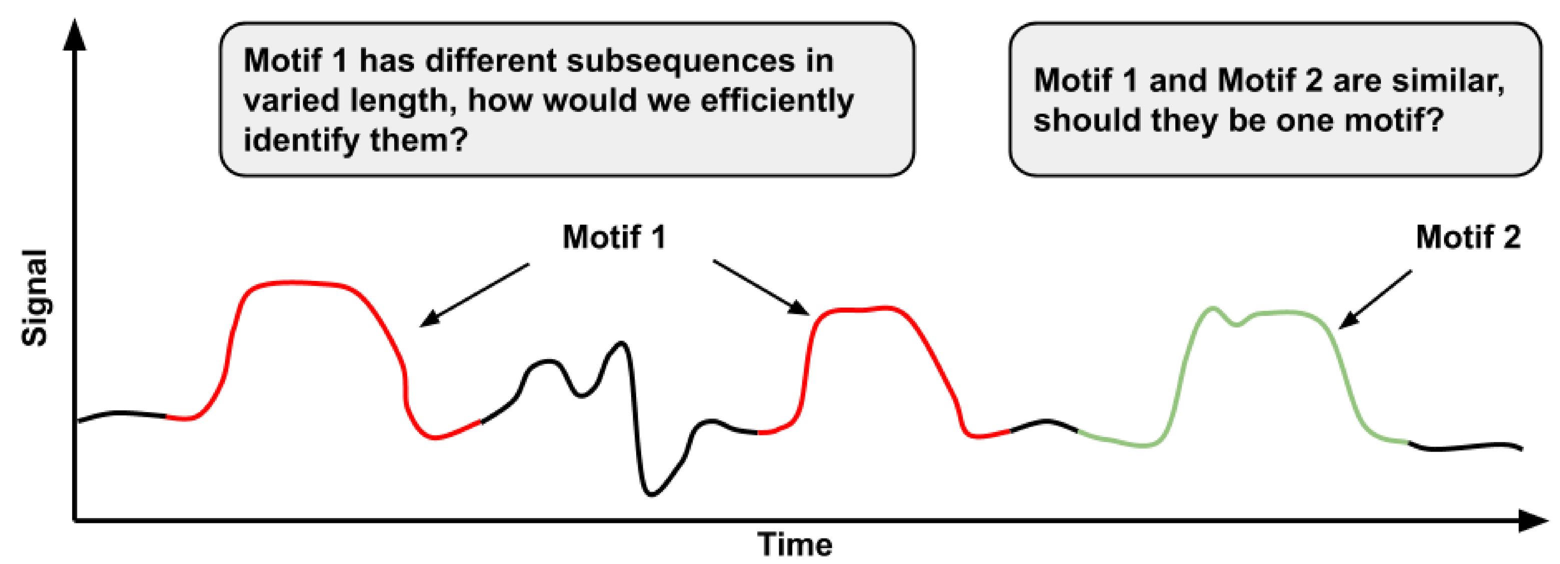

Figure 1: Challenges in motif discovering.

- Motif length could **vary** in wearable sensor time-series.
- How to identify the **optimal** set of motifs?

Contribution:

- A variable-length motif learning approach that combines SAX-based motif match algorithm and principle optimization.
- The proposed pipeline can capture useful structure for human behavioral analysis and modeling from heart rate data collected in the wild

Definition

- $(t_1, t_2, ..., t_i, t_{i+1}, ..., t_n) \Longrightarrow$ The time series data T of length n.
- $s_{p,q} \Longrightarrow$ A contiguous set of samples (subsequence) start from point p and end at q.
- $u = (u_1, u_2, ..., u_L) \Longrightarrow$ The piecewise aggregate approximation (PAA) representation of $s_{p,q}$.
- Word $w = w_1 w_2 ... w_L \Longrightarrow$ Symbolic representation of a subsequence.
- Top-K motifs $\Longrightarrow$ K Motifs that maximizes the total number of recurrent subsequences, while pairwise Euclidean distances between different motifs are above a threshold value.

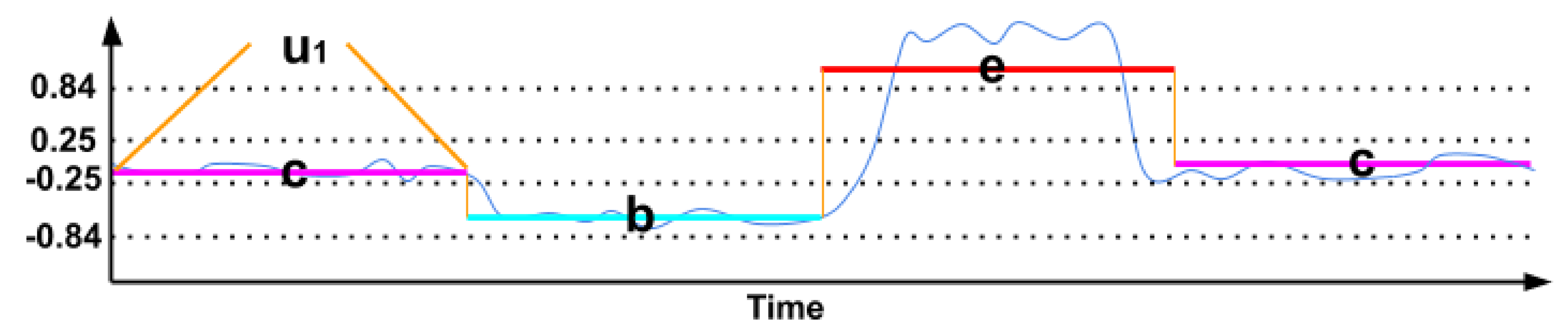

Figure 2: Example of Generating SAX word.

Method

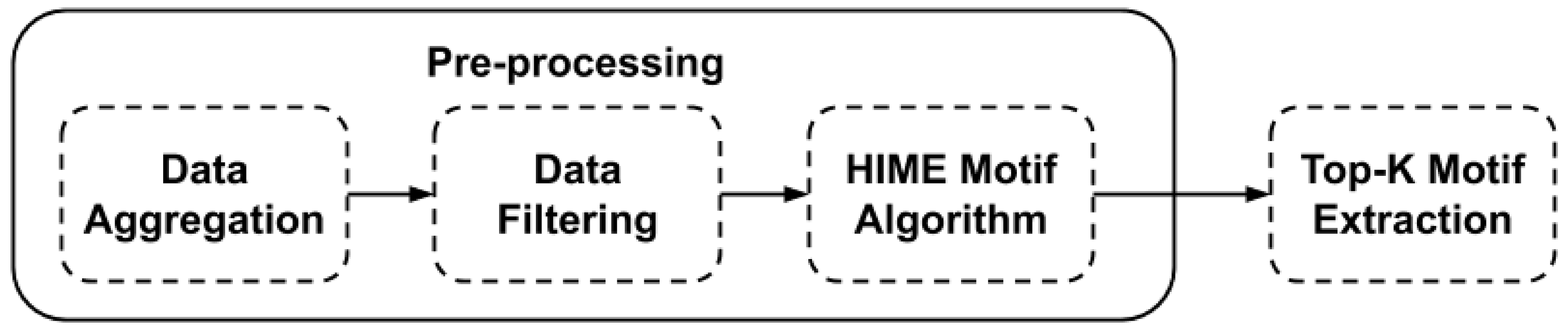

Figure 3: Optimal variable-length time-series motif learning pipeline.

Pre-processing

- **Data aggregation** $\Longrightarrow$ The time-series is aggregated every 3 minute.
- **Data filtering** $\Longrightarrow$ Savitzky-Golay filter.
- **HIME** (HierarchIcal based Motif Enumeration) $\Longrightarrow$ Variable-length motifs are detected using in a single SAX word $w = w_1 w_2 ... w_L$.

Subsequences in PAA	Subsequences in Symbols
(0.96, 0.71, 0.76, 0.31, 0.12, -0.23)	(a, a, a, b, b, c)
(1.12, 1.05, 0.85, 0.62, 0.57, 0.43)	(a, a, a, b, b, b)
(1.21, 1.11, 0.77, 0.42, 0.11, 0.35)	(a, a, a, b, b, b)
(-0.81, -0.62, -0.54, -0.31, 0.02, 0.12)	(d, c, c, c, b, b)
......	

Should they belong to the same motif?

Should we discard this rarely occurring motif?

Figure 4: How should we choose the top-K motifs?

Top-K Motif Learning

- Matrix $U \Longrightarrow$ PAA representation of J subsequences ($U \in \mathbb{R}^{J \times L}$), where each row in U represents a PAA vector of one subsequence of length L.
- Define total frequency given M with K motifs and distance threshold D:

$$F(M) = \sum_{k=1}^{K} \sum_{j=1}^{J} F_{k,j}, \text{ where } F_{k,j} = \begin{cases} 1, & \text{if } \sum_{l=1}^{L} (M_{k,l} - U_{j,l})^2 < D \\ 0, & \text{otherwise} \end{cases}$$

- We want function $F(M)$ is maximized while K motifs are different from each other by $2 \times D$:

$$M^* = \underset{M \in \mathbb{R}^{K \times L}}{\text{argmax}} F(M), \text{ subject to } \sum_{l=1}^{L} (M_{k,l} - M_{h,l})^2 > 2D$$
$$\text{where } k \in \{1, ..., K\}, \text{ and } h \in \{k+1, ..., K\}$$

- We approximate $F(M)$ using a Gaussian kernel to handle zero derivative of $F(M)$ and discontinuity at the point where $\sum_{l=1}^{L} (M_{k,l} - U_{j,l})^2 = 0$.
- Apply gradient ascent solution to solve this optimization problem.

TILES Dataset

- **TILES**, **Tracking IndividuaL pErformance with Sensors**, is a comprehensive human-subject experiments conducted in early 2018 to examine how the physiological, environmental, and behavioral variables impact job performance and employee wellness.
- The cohort used in this work includes over 100 individuals working as full-time nursing professions and 84 individuals worked the day shift
- This work focuses on analyzing PPG data collected using Fitbit Charge 2

	All Nurses	Day-shift	Night-shift
Average	1162.63	1182.50	1131.30
Standard Deviation	403.48	401.51	404.60

Table 1: Average/Std of valid PPG recording length in hours in TILES study.

Results

- Examples of learned optimal motifs.

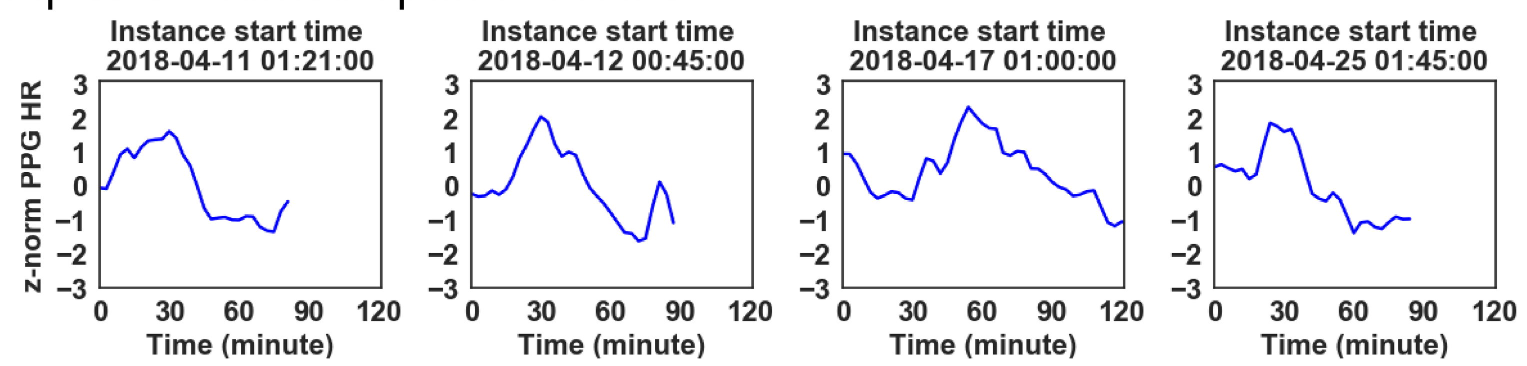

Figure 5: 4 subsequences of a motif that occurred repeatedly during the start of sleep

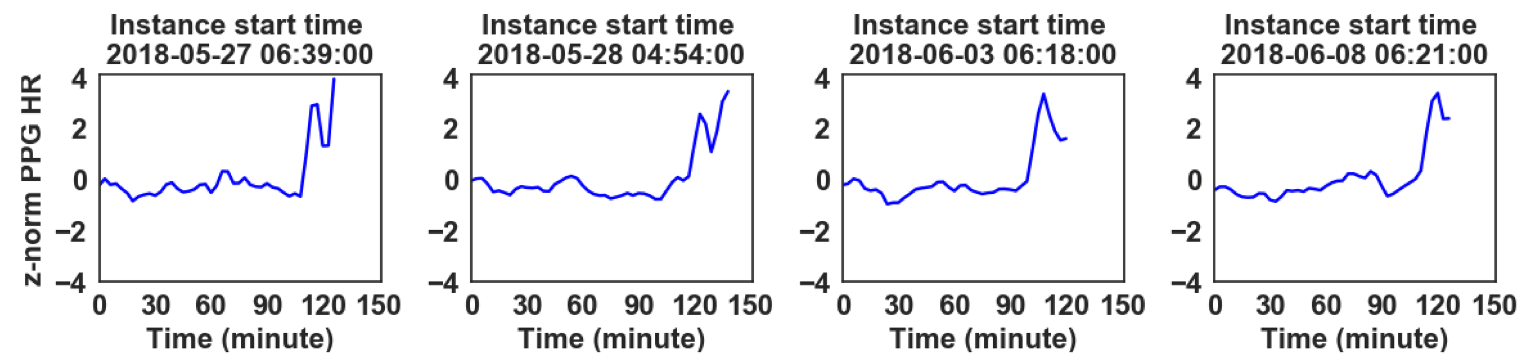

Figure 6: 4 subsequences of a motif that occurred repeatedly during the end of sleep

- Applying motif features to predict work status using motif frequencies from our pipeline and HIME algorithm.

	HIME Motif	Motif, D at pct=1%	Motif, D at pct=2%	Motif, D at pct=3%
Accuracy	65.20%	65.62%	67.20%	**68.54%**

Table 2: Prediction accuracy of work status using motif derived features

Acknowledgement

- Thanks for Intelligence Advanced Research Projects Activity (IARPA) and MOSAIC program for their support.