Spring 2024

Combine Shapelets

Zeng Qingwen
Loyola Marymount University, zqw931003@gmail.com

Combine Shapelets

by

Qingwen Zeng

A thesis presented to the

Faculty of the Department of
Computer Science
Loyola Marymount University

In partial fulfillment of the
Requirements for the Degree
Master of Science in Computer Science

April 19, 2024

This thesis has been examined and approved in partial fulfillment of the requirements for the Degree Master of Science in Computer Science by:

Thesis Advisor /Chair of Committee:

Delaram Yazdansepas, Ph.D.

Committee Members:

John David N. Dionisio, Ph.D.

Mandy B. Korpusik, Ph.D.

1 Abstract

Sensor-based human activity recognition has become an important research field within pervasive and ubiquitous computing. Techniques for recognizing atomic activities such as gestures or actions are mature for now, but complex activity recognition still remains a challenging issue[6]. I was a candidate in an activity classification thesis[3]. It collected 4 activities, which included walking on the sidewalk for a set distance, walking up and down a set of stairs, walking on the treadmill at 2.5 mph for 2 minutes, and jogging on the treadmill at 5.5 mph for 1 minute. It took 30 minutes to collect one candidate data. If complex activity data can be made up with atomic activities data, the data collecting process will be simplified. In this thesis, I used methods to mimic a complex activity shapelet by combing atomic activity shapelets. I first collect two candidates walk, jump and skip time series data, in which walk and jump are considered the atomic activities of skip. Time series patterns, shapelets, are extracted using tsshapelet package. Shapelets are small sub-series, or parts of the time-series, that are informative or discriminative for a certain class. They can be used to transform the time-series to features by calculating the distance for each of the time-series you want to classify to a shapelet[9]. In order to create skip representative shapelet, Barycenter Dynamic Time Warping and Weighted Dynamic Time Warping are used to average walk and jump shapelet, and then compare the euclidean distance between skip shapelet with walk shapelet, jump shapelet and, combined-shapelet. Experimental result show that the combined-shapelet is closer to skip shapelet than single walk or jump shapelet. Then I use three evaluation methods to mathematically and statistically show that combined-shapelet and real skip shapelet are similar. Evaluation methods include sliding window, cycle comparison and random comparison. To verify whether combined-shapelet can substitute real skip shapelet, a new labeled time series data is introduced, the result shows that both shapelets have the label accuracy around 70%, accuracy difference is less than 1%.

2 Methodology

2.1 Data

In the thesis, I focus on the relationships among jump, walk and skip. The assumption is that average of jump and walk shapelets can represent skip. Figure1, Figure 2 and Figure 3 are labeled jump, walk and skip accelerometer time series data. The data was collected from the accelerometer set at candidates' waist, because the location is close to the place we put the portable device[7], and the waist data won't be affected by domain leg problem. It shows that different individuals have distinct accelerometer time series data, because individuals have different body movement patterns, exercise habits, and so on. It also shows that the data for the same person engaged in different activities can vary due to the unique characteristics and motion patterns associated with each activity.

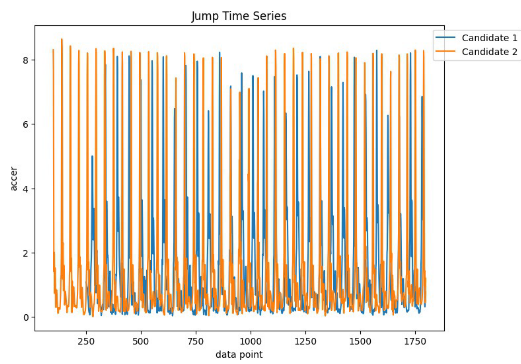

Figure 1: Candidates accelerometer Jump time series

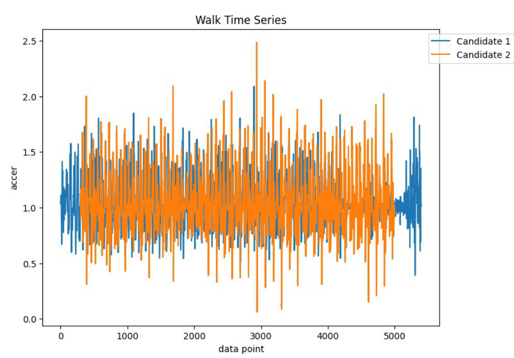

Figure 2: Candidates accelerometer Walk time series

2.2 Preprocess

As the user moves, the accelerometer stores the acceleration values associated to waist sensor. The accelerometer provides acceleration values of each of its three

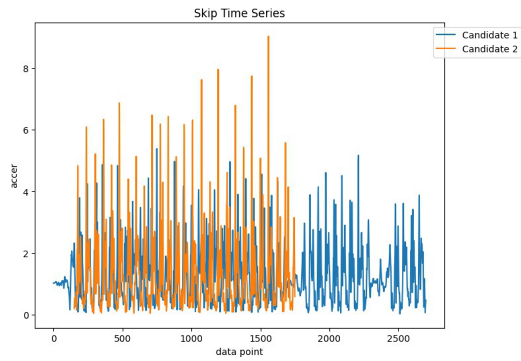

Figure 3: Candidates accelerometer Skip time series

axis. The position of the device and its orientation affects the data captured (the height, orientation and position on the body). In lab experiments this may not be so problematic. But, focusing in real case scenarios if the phone is placed in the pocket its center of axis may move[5]. In this same scenario, it cannot be ensured the exact position of the device. To solve this problem, it was decided to unify the three signals into a single one. The method more used to converge all signals is by the quadratic sum of the accelerations. By this, not only the orientation of the smartphone influence will be reduced, but also will improve the system performance related to resource usage and processing time[4].

$$R = \sqrt{(x^2 + y^2 + z^2)}$$

2.3 Cycle Extraction

Before process the time series data, I zoomed in the plot to find the patterns. There is a repetitive pattern in each activity, and the pattern is separated by the global maximum. For example, in Figure 4, there are 4 decremented local maximum within 2 global peak. In Figure 5, there are 3 local maximum, and second one is global peak. In Figure 6, the pattern is more complex, it needs methods introduced in Methodology part to extract. In [10], the author extracted activity shapelets by three steps: peak detection, shapelet extraction and shapelet selection. In [4], the author finds all local maximum points in the vector, using a small window. By doing so, several maximum points are found. Among these points there should be the starting points of gait cycle. But also included, there are several characteristic points of the user's gait as well as noise values[1]. Therefore, I use following steps

to eliminate those candidates that are not the starting points.
Set the minimum global maximum threshold and peak series between two global maximum as a cycle. There is a method used to choose the suitable threshold. It randomly set the threshold and calculate the result cycle length each time. And then pick the threshold with consistent cycle length, because if the threshold is too big or small, the cycle length distribution will be more sparse.

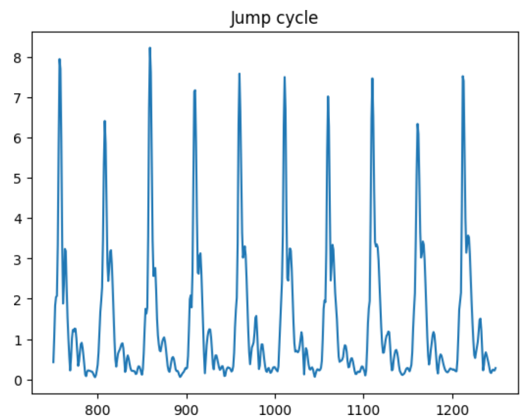

Figure 4: Raw Jump Cycle

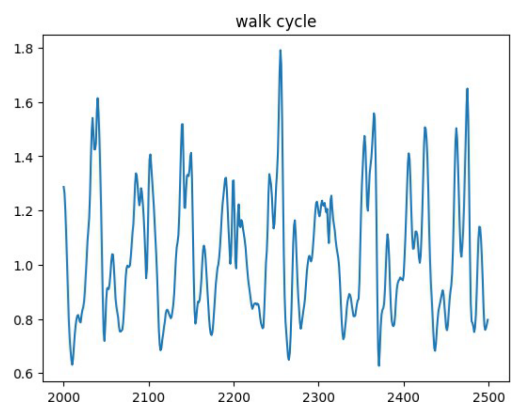

Figure 5: Raw Walk Cycle

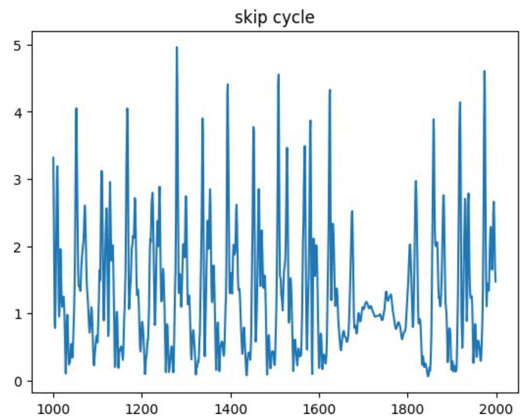

Figure 6: Raw Skip Cycle

2.4 Distance Calculation

In this thesis, dynamic time warping is used to calculate two shapelets distance. Dynamic time warping (DTW) is a way of comparing two, temporal sequences that don't perfectly sync up through mathematics[2]. DTW is widely used on sound pattern recognition and stock market prediction, which focus on up and down in a series, and accelerometer time series data is also with various rise and fall, so I use DTW to calculate two shapelet difference. The reason I used DTW over Euclidean distance to calculate distance is because that DTW can capture temporal variations, while euclidean distance can't. The reason is that DTW considers different possible alignments and chooses the one that minimizes the total cost, but euclidean distance assumes a one-to-one mapping between points in two sequences. Considering that the large size of time series data and regular DTW comes with a high computational cost, I used FastDTW, an approximation of DTW that has a linear time and space complexity that uses a multilevel approach. It avoids the brute-force dynamic programming approach of the standard DTW algorithm by using a multilevel approach. The time series are initially sampled down to a very low resolution. A warp path is found for the lowest resolution and "projected" onto an incrementally higher resolution time series. The projected warp path is refined and projected again to yet a higher resolution. The process of refining and projecting is continued until a warp path is found for the full resolution time series[8].

3 Combine Shapelet

The jump, walk and skip representative shapelets are in Figure7. In order to find an appropriate way to create a skip shapelet base on jump and walk shapelets, It is important to observe shapelets features and make assumptions. It shows that the wave frequency of skip shapelet is similar to walk shapelet, because their left right switch feature. And the peak value of skip shapelet is about half of the jump shapelet. In Figure7 shows that the length of walk and skip shapelets are about 100, double the value of jump shapelet, because it takes one left and one right as a cycle. For this reason, I use two consecutive jump shapelet as jump shapelet in the combining process. It is used as Figure8.

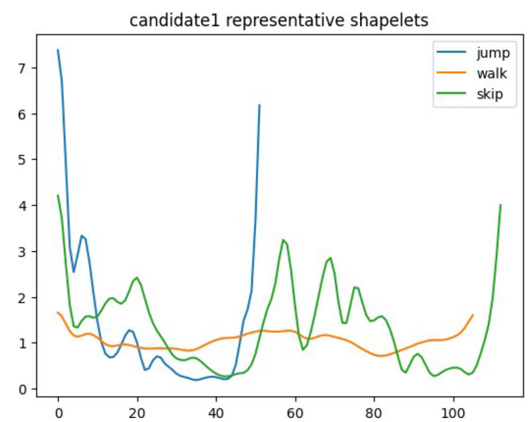

Figure 7: Candidate 1 representative shapelets

3.1 Soft Dynamic Time Warping Barycenter Average

I first assume that combine the geometric similarity of jump shapelet and walk shapelet can get skip shapelet, so I want to use Dynamic Time Warping related average method. The reason is that DTW is a measure of similarity between two sequences that accounts for the phase shifts and warping in the time domain. I used Soft Dynamic Time Warping Barycenter Average, which is a custom version of Dynamic Time Warping Barycenter Average, and it is allowed to change weights. The procedure of Dynamic Time Warping Barycenter Average is involves aligning time series using DTW and iterative updating the barycenter. The result of average is in Figure9. It shows that the waves in walk shapelet are hid-

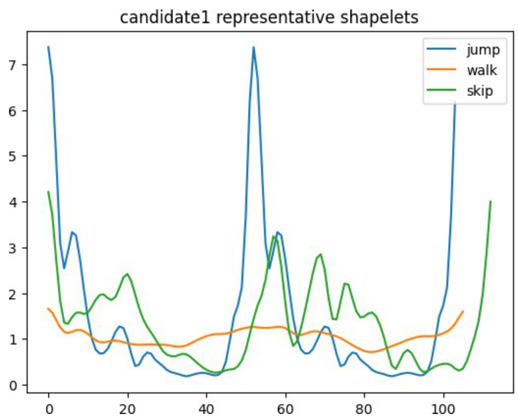

Figure 8: Candidate 1 Processed Jump Shapelets

den, because they are averaged by jump shapelets. I concluded that average the up and falls in walk and jump shapelet won't get skip shapelet.

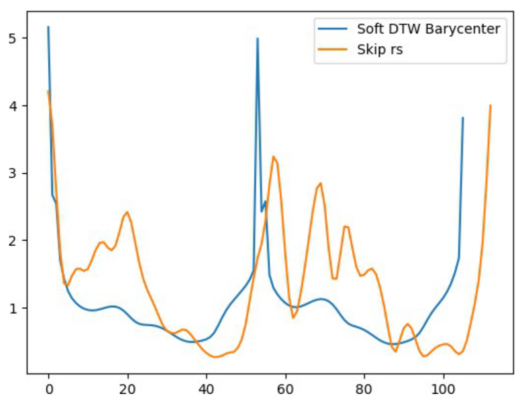

Figure 9: Candidate 1 Soft DTW Barycenter

3.2 Euclidean Barycenter

Figure7 shows that the length of walk and skip shapelets length are close, I think that I need to find a way that focus on the sequence of time series. I used euclidean barycenter. It is a method to average points in sequence. The result is in Figure10, the shapelets in this plot look similar, so I will evaluate it accuracy on labeled skip time series and an unlabeled time series data.

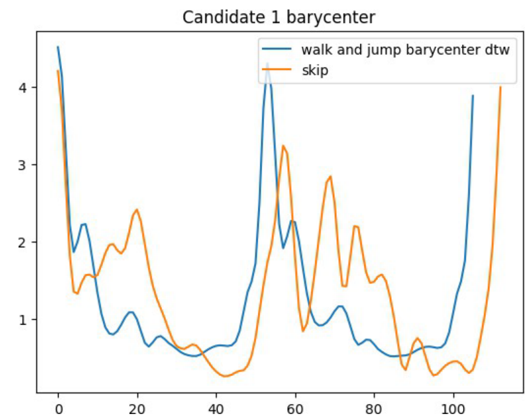

Figure 10: Candidate 1 Euclidean Barycenter Shapelet

3.3 Weighted Barycenter

Peaks in Figure10 are not perfect match, so I assume to add weight index in euclidean barycenter. In order to find the best weight, I iterative calculate the DTW distance of created shapelet and skip shapelet, and then pick the index with lowest distance. The result is index of walk is 0.5, and index of jump is 1. The result is in Figure11.

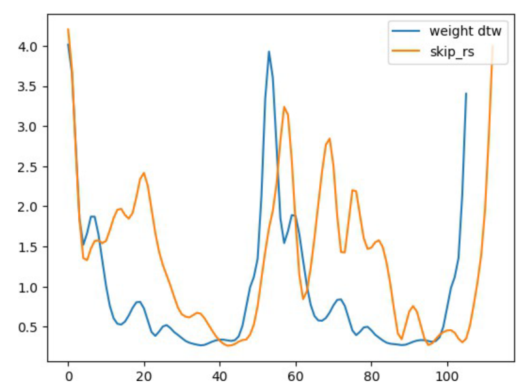

Figure 11: Candidate 1 Weighted Barycenter Shapelet

4 Shapelet Evaluation

4.1 Labeled Dataset

A new evenly distributed time series data is collected, which include equal length of walk, jump and skip activities, and each activity switch point is recorded. Using both sliding window and cycles comparison to evaluate the real skip shapelet and combined shapelet. For each data point, calculate the distance between a window and three shapelets in dictionary respectively, and then label the point activity with the shortest distance shapelet. For example, if at data point A, the DTW distance to walk smaller than jump and skip, label the point A walk. After sampling and calculation, count the correct label accuracy.

4.2 Sliding Window

The sliding window technique involves moving a fixed-size window through the time series and performing computations within that window. After getting the combined shapelet and skip shapelet, calculate the distance between proposed shapelet and labeled skip time series by sliding window, of which window size is same as shapelet length. If the combined shapelet can replace skip shapelet, their distance result should be close.

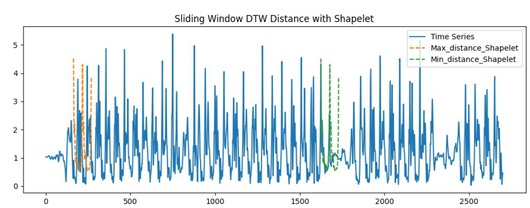

Figure 12: Barycenter of walk and jump shapelets

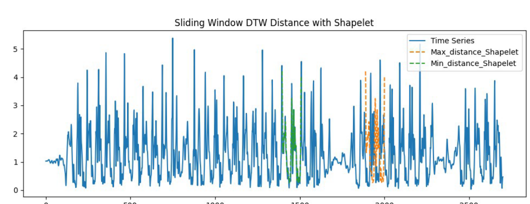

Figure 13: Skip representative shapelet

4.3 Cycles

Activity time series exhibits periodic patterns, so cutting it into cycles is better for finding underlying relationships between time series data and shapelet. A labeled skip time series data is cut into cycles, a series between two global maximum is a cycle, and then store all cycles in a cycle set. Calculate the DTW distance between each cycles with combined shapelet and real skip shapelet respectively.

4.4 Random Sampling

In the real scenario, an activity time series is a large dataset. Use Sliding window to whole sequence requires large computational resource and time consuming. Random sampling reduces the number of computations compared to analyzing the entire sliding window. In this thesis, the result shows that random sample half of the dataset has same accuracy as apply whole time series.

5 Conclusion

The result of sliding combined shapelet and skip shapelet on a skip time series is in Figure12, Figure13, Table1 and Table2. Min, Max and Mean in Table1 and Table2 are the dynamic time warping distance of combined shapelet and real skip shapelet from a part of skip time series. The distance values are similar, but sliding window prediction accuracy of both shapelets on a new labeled data is around 50%, which is same as random choice. In order to find the reason, I compare the real label with guessed label, it shows that bad accuracy is from the dominant of jump shapelet. In Figure7, jump shapelet looks flat and lack of information, so I changed the jump shapelet to Figure15, which includes two up and down. After the change, the sliding window accuracy is enhanced to 60%. For cycle extraction is 71%, in which walk is 100%, jump is 70%, skip is 61%. It is still not a good accuracy, there are three possible reasons. Firstly, the combining process is based on the guess that skipping is an activity that walking and jumping happens concurrently, but it might be an activity that walking and jumping happens both concurrently and sequentially. Secondly, inter-relationships are not considered, but that are important for understanding the activity semantics. At last, as skipping is a complex activity, one shapelet may not enough to represent its features. The result showed that simple euclidean barycenter and dtw barycenter are not complex enough to get all the skipping features from walking and jumping activities, and the boundary between skipping and jumping is not clear. This is because of three features of complex activity.

- Difficulty Handling Variable-Length Sequences: Complex activities may exhibit variability in duration, making it challenging for shapelet-based models that assume fixed-length sub-sequences.
- Lack of Global Context: Shapelets typically focus on local patterns within the time series, while they might miss capturing the global context of complex activities. Global context is crucial for understanding the overall structure and semantics of complex behaviors.
- Sensitivity to Noise: Shapelets are sensitive to noise and variations in the data. Because more movement in complex behaviour than simple activity, too much noise might affect the shapelet performance, which leading to decreased classification accuracy.

6 Future Work

In this thesis, there are only three known activities, jump, walk and skip. In the future, if more different activities data are collected, I can label the data and extract features from it, and then add the feature to a decision tree. In this way, if an unknown activity time series comes, the model can label it as None. And, I can collect more data to find best weight index of weighted barycenter. For now, shapelet extraction and combining treated each data points independently without considering inter-relationships, in the future, I will handle temporal relationships between activities.

Candidate 1	Mean	Min	Max
Combined_shapelet	39.411953	21.896391	65.332765
Skip_shapelet	33.027806	14.972689	63.371953

Table 1: Candidate 1 DTW Distance

Candidate 2	Mean	Min	Max
Combined_shapelet	50.725114	27.084183	66.0174
Skip_shapelet	42.372542	20.139492	59.80756

Table 2: Candidate 2 DTW Distance

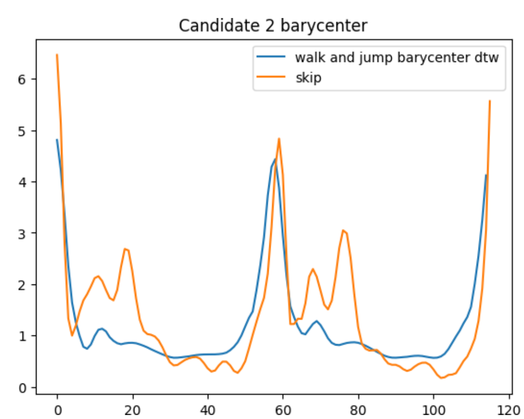

Figure 14: Candidate 2 Euclidean Barycenter Shapelet

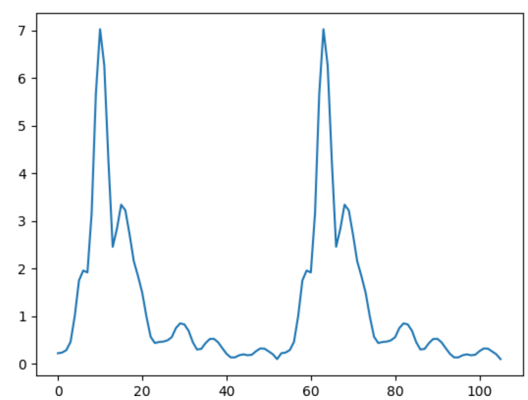

Figure 15: New Jump Shapelet

References

[1] Heikki Ailisto, Mikko Lindholm, Jani Mäntyjärvi, Elena Vildjiounaite, and Satu-Marja Mäkelä. Identifying people from gait pattern with accelerometers. *Proc SPIE*, 5779, 03 2005.

[2] Esmaeil Alizadeh. An introduction to dynamic time warping. In *Buitin*, 2022.

[3] Grant Ellison, Milla Penelope Markovic, and Delaram Yazdansepas. Real-time human activity classification using gait cycle averaging and biometric heuristics. In *2023 International Conference on Machine Learning and Applications (ICMLA)*, pages 355–361, 2023.

[4] Pablo Fernandez-Lopez, Judith Liu-Jimenez, Carlos Sanchez-Redondo, and Raul Sanchez-Reillo. Gait recognition using smartphone. In *2016 IEEE International Carnahan Conference on Security Technology (ICCST)*, pages 1–7, 2016.

[5] D. Gafurov, K. Helkala, and T. Soendrol. Gait recognition using acceleration from mems. In *First International Conference on Availability, Reliability and Security (ARES'06)*, pages 6 pp.–439, 2006.

[6] Li Liu, Yuxin Peng, Shu Wang, Ming Liu, and Zigang Huang. Complex activity recognition using time series pattern dictionary learned from ubiquitous sensors. *Information Sciences*, 340-341:41–57, 2016.

[7] Jani Mäntyjärvi, Mikko Lindholm, Elena Vildjiounaite, Satu-Marja Mäkelä, and Heikki Ailisto. Ailisto, "identifying users of portable devices from gait pattern with accelerometers. volume 2, pages ii/973 – ii/976 Vol. 2, 04 2005.

[8] Stan Salvador and Philip K. Chan. Fastdtw: Toward accurate dynamic time warping in linear time and space. 2004.

[9] Gilles Vandewiele. Enriching shapelets with positional information for time-series classification. In *Towards Data Science*, 2020.

[10] Delaram Yazdansepas, Nitin Saroha, Lakshmish Ramaswamy, and Khaled Rasheed. Towards efficient real-time human activity recognition using wearable sensors: A shapelet-based pattern matching approach. 10 2018.