The Journal of Open Source Software

seg1d: A Python package for Automated segmentation of one-dimensional (1D) data

Mathew Schwartz[1], **Todd C. Pataky**[2], **and Cyril J. Donnelly**[3]

1 New Jersey Institute of Technology **2** Kyoto University, Department of Human Health Sciences **3** Nanyang Technological University, Rehabilitation Research Institute of Singapore

DOI: 10.21105/joss.02404

Software

- Review
- Repository
- Archive

Editor: Tania Allard

Reviewers:

- @AKuederle
- @ejhigson

Submitted: 20 May 2020
Published: 11 August 2020

Summary

The term segmentation refers to the division of a data series into segments. In this paper, segments refer to a contiguous time series which is a subset of another time series (Bouchard, 2006). Identifying these segments requires a reference series to be found in the larger series, which can be an exact copy or an approximation. When an approximation of the desired segment is used, the metric and method for identifying similarity of all subsets is essential to achieving a desired result. Additionally, multiple references may be desired to further describe the aspects of segment similarity, some of which may be more important than others.

Segmenting human motion is a topic studied in various fields such as robotics (e.g., humanoids), biomechanics, and computer graphics (e.g., gaming and animation). In the human movement sciences, segmentation is often performed as the labeling of specific events, such as the components of a gait cycle, for analysis and interpretation of the data within these labeled bounds. Subjectively defining what is a movement or a phase of a movement can be particularly difficult due to variations in what one may define as a single movement. As such, the points at which the movement or phases of a movement starts and ends can be ambiguous. The averaging of multiple features (e.g., marker trajectories, joint angles, or other information derived from the data) of a movement or even multiple movements (e.g., multiple marker trajectories from multiple observation sets) allows for a tolerance to some individual features failing to provide an expected characteristic (e.g., a signal above or below a threshold value from a force plate) that may normally be relied upon for identifying an event. Defining weights of individual features, with either algorithmic approaches or through expert knowledge, further facilitates segmentation of similar movements.

With feature-rich data such as multiple marker trajectories from motion capture, the reduction of meaningful features is important for segmentation performance (Bouchard & Badler, 2015). The use of marker trajectory and ground reaction force, without computing kinematics, has been shown to be sufficient in movement segmentation tasks (Lin, Bonnet, Joukov, Venture, & Kulic, 2016). Unique feature creation, extraction, and storage can be used to additionally index databases for fast movement-based time-series data retrieval (Kapadia, Chiang, Thomas, Badler, & Kider Jr, 2013). Subseries searching of databases has often been performed with similarity metrics, each with their own individual downfalls, e.g.: Longest Common Subseries (LCSS), Euclidean Distance (ED), and Dynamic Time Warping (DTW) (Vlachos, Hadjieleftheriou, Gunopulos, & Keogh, 2003). Discrete Fourier Transformation (DFT) has also been used as a method for improving the efficiency of windowed correleations (Zhu & Shasha, 2002). Data reduction techniques for optimization can also be used in windowed correlations of generic time-series data (Cole, Shasha, & Zhao, 2005).

Peak detection has been used to identify gait events through the inference of a cyclic motion and reducing reliance on physical meanings of the signal by searching for the assumed cyclic

pattern rather than a given threshold value or calculated joint angle (Jiang, Wang, Kyrarini, & Gräser, 2017). Peak detection from cross-correlated data has been used for gait event detection in accelerometry data (Yoneyama, Kurihara, Watanabe, & Mitoma, 2013). Alternative methods for segmenting data have used physical devices to act as switches on foot contact (Agostini, Balestra, & Knaflitz, 2013). Most recently, deep neural networks have been used to predict foot contact, but requires a large amount (i.e., thousands of trials) of training data (Kidziński, Delp, & Schwartz, 2019). Using DTW, (Sarsfield et al., 2019) was able to identify movements in realtime with a single reference segment. Allowing both single and multiple reference segments, as well as multiple features and optional weights, the segmentation of various data is more practical in a single package.

Statement of Need

Subsequence identification and similarity between a reference(s) and target data items is a commonly desired task done both manually and automatically in a variety of fields. The ability to further automate and create reliable, consistent results, is of importance for many data processing related tasks. For example, in typical motion capture sessions of walking gait in a lab, embedded force plates provide high fidelity measurements for foot-strike and toe-off. However the cycles before and after this event are often discarded. By using a collection of features from a known segment (i.e., the cycle over the force-plate), similar sequences within a trial can be found and used for study. Furthermore, some movements are not so well defined with these external sensing tools, and rather a template movement selected by a human is the most reasonable way to identify the sequence of data describing a particular motion.

seg1d is an open-source Python package for the automated segmentation and extraction of time series data using one or more reference sequences. The segmentation process allows users to apply various methods and parameters for the process through weighted reference features in a rolling correlation size-varying window of any scale below the length of the targeted data. Correlations can be averaged across the references and a peak detection algorithm finds individual segments. Non-overlapping segments are identified and a clustering algorithm groups the most similar subsequence movements within the target. The package was developed for movement sciences but can be useful to anyone interested in extracting correlated subsequences from a dataset.

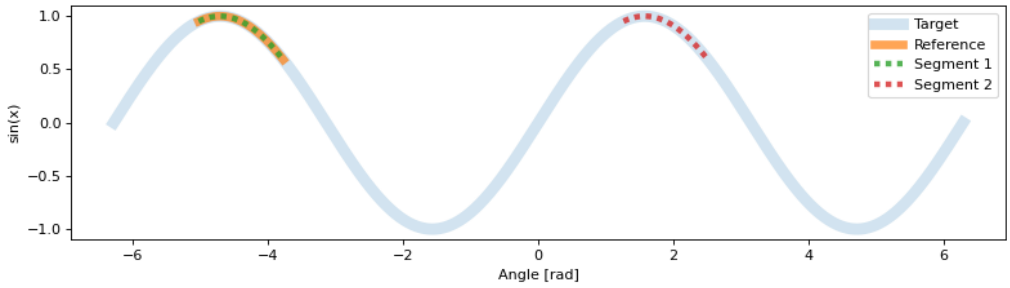

Figure 1: Sample segments in a timeseries from a reference

Acknowledgements

This work was supported in part by the Agency for Science, Technology and Research (A*STAR), Nanyang Technological University (NTU) and the National Health Group (NHG) (RRG3: 2019/19002).

Todd Pataky is supported through the Kiban B Grant 17H02151 (Japan Society for the Promotion of Science).

References

Agostini, V., Balestra, G., & Knaflitz, M. (2013). Segmentation and classification of gait cycles. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, *22*(5), 946–952. doi:10.1109/tnsre.2013.2291907

Bouchard, D. (2006). Automated time series segmentation for human motion analysis. *Center for Human Modeling and Simulation, University of Pennsylvania*.

Bouchard, D., & Badler, N. I. (2015). Segmenting motion capture data using a qualitative analysis. In *Proceedings of the 8th acm siggraph conference on motion in games* (pp. 23–30). ACM. doi:10.1145/2822013.2822039

Cole, R., Shasha, D., & Zhao, X. (2005). Fast window correlations over uncooperative time series. In *Proceedings of the eleventh acm sigkdd international conference on knowledge discovery in data mining* (pp. 743–749). doi:10.1145/1081870.1081966

Jiang, S., Wang, X., Kyrarini, M., & Gräser, A. (2017). A robust algorithm for gait cycle segmentation. In *2017 25th european signal processing conference (eusipco)* (pp. 31–35). IEEE. doi:10.23919/eusipco.2017.8081163

Kapadia, M., Chiang, I.-k., Thomas, T., Badler, N. I., & Kider Jr, J. T. (2013). Efficient motion retrieval in large motion databases. In *Proceedings of the acm siggraph symposium on interactive 3D graphics and games* (pp. 19–28). doi:10.1145/2448196.2448199

Kidziński, Ł., Delp, S., & Schwartz, M. (2019). Automatic real-time gait event detection in children using deep neural networks. *PloS one*, *14*(1), e0211466. doi:10.1371/journal.pone.0211466

Lin, J. F.-S., Bonnet, V., Joukov, V., Venture, G., & Kulic, D. (2016). Comparison of kinematic and dynamic sensor modalities and derived features for human motion segmentation. In *2016 ieee healthcare innovation point-of-care technologies conference (hi-poct)* (pp. 109–112). IEEE. doi:10.1109/hic.2016.7797709

Sarsfield, J., Brown, D., Sherkat, N., Langensiepen, C., Lewis, J., Taheri, M., Selwood, L., et al. (2019). Segmentation of exercise repetitions enabling real-time patient analysis and feedback using a single exemplar. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, *27*(5), 1004–1019. doi:10.1109/tnsre.2019.2907483

Vlachos, M., Hadjieleftheriou, M., Gunopulos, D., & Keogh, E. (2003). Indexing multidimensional time-series with support for multiple distance measures. In *Proceedings of the ninth acm sigkdd international conference on knowledge discovery and data mining* (pp. 216–225). doi:10.1145/956750.956777

Yoneyama, M., Kurihara, Y., Watanabe, K., & Mitoma, H. (2013). Accelerometry-based gait analysis and its application to parkinson's disease assessment—part 1: Detection of stride event. *IEEE Transactions on neural systems and rehabilitation engineering*, *22*(3), 613–622. doi:10.1109/tnsre.2013.2260561

Zhu, Y., & Shasha, D. (2002). Statstream: Statistical monitoring of thousands of data streams in real time. In *VLDB'02: Proceedings of the 28th international conference on very large databases* (pp. 358–369). Elsevier.