



PSEUDO: Interactive Pattern Search in Multivariate Time Series with Locality-Sensitive Hashing and Relevance Feedback

Y. Yu^{1,2} , D. Kruyff¹, J. Jiao^{1,3}, T. Becker² and M. Behrisch¹ 

¹Utrecht University, Netherlands

²IAV GmbH Ingenieurgesellschaft Auto und Verkehr, Germany

³Fraunhofer Institute for Systems and Innovation, Germany

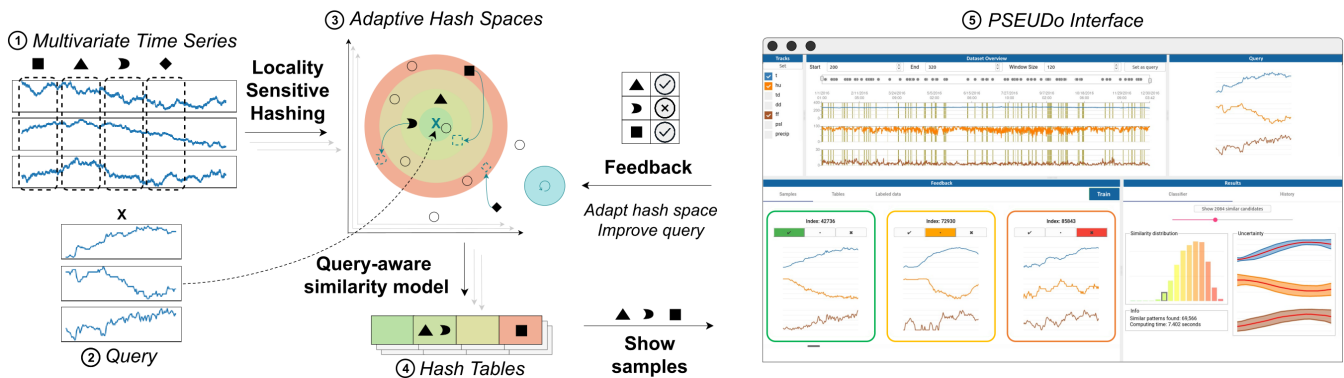


Figure 1: PSEUDO creates a representation model for multivariate time series based on locality-sensitive hashing, conducts scalable pattern retrieval with few initial labels and evolves with an interpretable relevance feedback mechanism to capture subjective pattern similarity.

Abstract

We present PSEUDO, a visual pattern retrieval tool for multivariate time series. It aims to overcome the uneconomic (re-)training with deep learning-based methods. Very high-dimensional time series emerge on an unprecedented scale due to increasing sensor usage and data storage. Visual pattern search is one of the most frequent tasks on such data. Automatic pattern retrieval methods often suffer from inefficient training, a lack of ground truth, and a discrepancy between the similarity perceived by the algorithm and the user. Our proposal is based on a query-aware locality-sensitive hashing technique to create a representation of multivariate time series windows. It features sub-linear training and inference time with respect to data dimensions. This performance gain allows an instantaneous relevance-feedback-driven adaption and converges to users' similarity notion. We are benchmarking PSEUDO in accuracy and speed with representative and state-of-the-art methods, evaluating its steerability through simulated user behavior, and designing expert studies to test PSEUDO's usability.

CCS Concepts

• **Mathematics of computing** → Time series analysis; • **Information systems** → Users and interactive retrieval;

1. Introduction

Searching for patterns similar to a given query in a time series database is one of the most frequent problems in time series analysis [LKWL07]. In the literature, it is called pattern search [JDDH19, LPH*20], time series indexing [CKMP02], similarity search [NB05, GA16], query by content, sub-sequence

matching [FRM94], and twin search [GDK*21]. It is an abstraction of many real-world problems in natural science [LLS*20, LPH*20], engineering [LL18], medicine [JDDH19, GDK*21], and economics [NB05, TFRC07]. It remains an interesting and important question to efficiently and accurately search for patterns in unlabeled multivariate time series. Our automotive calibration engineers search for patterns spontaneously in measurement from en-

engine control units with over 10,000 tracks and wish for an answer as promptly as possible. This is challenging, not only because of the high dimensions (a large number of tracks), meager labels, and the efficiency requirement, but also the subjective and use-case-dependent similarity notion. Whereas model-free similarity measures lack trainable parameters and the power to model potentially complex similarity rules catering to the user's similarity notion, machine learning may suffer from few labels and slow training. Furthermore, our application engineers ask for an interpretable process, for instance, which tracks count most for the event behind the pattern, to assist the subsequent domain-specific analysis.

We propose PSEUDO, a tool for visual pattern retrieval in multivariate time series, especially very high-dimensional time series. It is powered by Locality-Sensitive Hashing (LSH) for multivariate time series [YLC*19]. In a nutshell, LSH linearly maps all tracks into one with groups of hash functions, making subsequent processing scalable with respect to the data dimensions. Our major contribution is the extension of this algorithm with an efficient, steerable, and interpretable relevance feedback mechanism. Relevance feedback is also introduced for tabular [BKSS14], text [vdBS*21] and image data [DPL*19]. It is first introduced to time series in [KP98] and appears recently in [LPH*20]. Finally, we implemented a user interface to assist the algorithm. Such UIs for time series retrieval are often called Visual Query Systems (VQSs) [LLS*20, SLW*20, LPH*20].

2. Implementation

In this work, we address the problem: how to conduct pattern search in very high-dimensional time series with very few ground truth labels efficiently for user interaction, accurately regarding the subjective similarity notion, and interpretably.

Our baseline is the Query-Aware Locality Sensitive Hashing (QALSH) algorithm [YLC*19]. We choose it as the core algorithm because of its speed and scalability for very high-dimensional time series. We extend the algorithm with an also efficient and explainable relevance feedback mechanism. The overall pipeline works as follows: 1) preprocessing with sliding windows and window normalization (① in Figure 1); 2) marking a pattern in the time series as the query by the user (②); 3) initial search with LSH (from ①② to ④); 4) sampling results for relevance feedback (from ④ to ⑤); 5) inspecting results and provide relevance feedback by the user (⑤); 6) evaluating feature/track importance, updating the LSH model and rerunning search (from ⑤ back to ③); 7) iterating the steps 4) to 6) until the user is satisfied with the result.

This query definition approach is called query-by-example [HB04, LPH*20]. A popular alternative is query-by-sketch [CG16, MA18, LLS*20]. We favor the former because the query can be unclear or overly complex for the user to draw.

To update the LSH model, we assume that the randomly initialized parameters of the hash functions are trainable. Moreover, we interpret them as feature/track importance. On the other hand, we can infer track importance by calculating and comparing the variances of tracks distances among the positive user labels and the query. The higher the variance, the lower the importance. This importance information extracted from the user labels is led back to

the parameters in the hash functions. We opt for this approach because it is very fast without mathematical optimization and is interpretable. The windows shown to the user for relevance feedback are sampled from both confident and unsure windows to avoid converging into a biased similarity notion. We also allow updating the query by averaging the query and the user labels with Dynamic Time Warping Barycenter Averaging (DBA) [PKG11].

We have implemented a prototypical UI to assist the algorithm. It has a main view showing tracks as line charts, a mini-map / range-slider with events as differently colored dots, a view to show the current query, a panel to show/hide tracks, a view to list sampled predictions and accept relevance feedback, a view to review the labels, a view for result statistics, and a view for training state management. For further interpretability, we visualize the classifiers / hash functions by calculating the mean shape and standard deviation of the windows perceived as similar by every single hash function in a dedicated view. Details can be found in Appendix A.

3. Results

We are evaluating the accuracy with and without relevance feedback and benchmarking speed (especially scalability with respect to dimensions) with quantitative experiments. On the other hand, we are designing expert studies to examine PSEUDO's usability.

Our first results look promising. The accuracy benchmark suggests that different datasets favor different methods, confirming the finding in [CG16]. LSH did not cause much accuracy loss. In the speed benchmark, LSH tightly followed the fastest similarity search algorithm Mueen's Algorithm for Similarity Search (MASS) for univariate time series and overtook the latter in multivariate cases. The scalability test with increasing dimensions also verified this finding: the speed advantage of LSH became more evident with more dimensions. In the experiment for relevance feedback, we witnessed improved accuracy with feedback rounds. We would like to inspect the track importance implied by the hash functions to check whether the relevance feedback mechanism has the desired feature selection effect.

4. Conclusion and Future Work

In this work, we propose PSEUDO, an adaptive and interpretable tool for pattern search in multivariate time series based on LSH and relevance feedback. It is particularly efficient for very high-dimensional time series and in use cases where initial labels are meager, and the promptness of the result counts. In the future, we expect an increasing collaboration between hashing algorithms and machine learning due to the explosion of data size, e.g., for massive video processing. This work is ongoing. We plan to examine scalable visualizations for high-dimensional data, such as [KKA95, HKA09, ATTA19] to better cooperate with the algorithm. We also need rigorous expert studies to evaluate PSEUDO's usability. Finally, we are working on relevance feedback beyond binary labels and visualization measuring quality of time series retrieval.

References

- [ATTA19] A. GOGOLOU, T. TSANDILAS, T. PALPANAS, A. BEZIRIANOS: Comparing similarity perception in time series visualizations. *IEEE Transactions on Visualization and Computer Graphics* 25, 1 (2019), 523–533. doi:10.1109/TVCG.2018.2865077. 2
- [BKSS14] BEHRISCH M., KORKMAZ F., SHAO L., SCHRECK T.: Feedback-driven interactive exploration of large multidimensional data supported by visual classifier. In *2014 IEEE Conference on Visual Analytics Science and Technology (VAST)* (2014), pp. 43–52. doi:10.1109/VAST.2014.7042480. 2
- [CG16] CORRELL M., GLEICHER M.: The semantics of sketch: Flexibility in visual query systems for time series data. In *2016 IEEE Conference on Visual Analytics Science and Technology* (2016), pp. 131–140. 2
- [CKMP02] CHAKRABARTI K., KEOGH E., MEHROTRA S., PAZZANI M.: Locally adaptive dimensionality reduction for indexing large time series databases. *ACM Trans. Database Syst.* 27, 2 (2002), 188–228. doi:10.1145/568518.568520. 1
- [DPL*19] DENNIG F. L., POLK T., LIN Z., SCHRECK T., PFISTER H., BEHRISCH M.: Fdive: Learning relevance models using pattern-based similarity measures. In *2019 IEEE Conference on Visual Analytics Science and Technology (VAST)* (2019), IEEE, pp. 69–80. 2
- [FRM94] FALOUTSOS C., RANGANATHAN M., MANOLOPOULOS Y.: Fast subsequence matching in time-series databases. In *Proceedings of 1994 ACM SIGMOD International Conference on Management of Data* (New York, NY, USA, 1994), SIGMOD '94, Association for Computing Machinery, pp. 419–429. doi:10.1145/191839.191925. 1
- [GA16] GIAO B. C., ANH D. T.: Similarity search for numerous patterns over multiple time series streams under dynamic time warping which supports data normalization. *Vietnam Journal of Computer Science* 3, 3 (2016), 181–196. doi:10.1007/s40595-016-0062-4. 1
- [GDK*21] GEORGIOS CHATZIGEORGAKIDIS, DIMITRIOS SKOUTAS, KOSTAS PATROUMPAS, THEMIS PALPANAS, SPIROS ATHANASIOU, SPIROS SKIADOPOULOS: Twin subsequence search in time series. *CoRR abs/2104.06874* (2021). 1
- [HB04] HARRY HOCHHEISER, BEN SHNEIDERMAN: Dynamic query tools for time series data sets: Timebox widgets for interactive exploration. *Information Visualization* 3, 1 (2004), 1–18. doi:10.1057/palgrave.ivs.9500061. 2
- [HKA09] HEER J., KONG N., AGRAWALA M.: Sizing the horizon: The effects of chart size and layering on the graphical perception of time series visualizations. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (New York, NY, USA, 2009), CHI '09, Association for Computing Machinery, pp. 1303–1312. doi:10.1145/1518701.1518897. 2
- [JDDH19] JOÃO RODRIGUES, DUARTE FOLGADO, DAVID BELO, HUGO GAMBOA: Ssts: A syntactic tool for pattern search on time series. *Information Processing & Management* 56, 1 (2019), 61–76. URL: <http://www.sciencedirect.com/science/article/pii/S0306457318302577>, doi:10.1016/j.ipm.2018.09.001. 1
- [KKA95] KEIM D. A., KRIEGEL H.-P., ANKERST M.: Recursive pattern: a technique for visualizing very large amounts of data. In *Proceedings Visualization '95* (1995), pp. 279–286. doi:10.1109/VISUAL.1995.485140. 2
- [KP98] KEOGH E. J., PAZZANI M. J.: An enhanced representation of time series which allows fast and accurate classification, clustering and relevance feedback. In *Kdd* (1998), vol. 98, pp. 239–243. 2
- [LKWL07] LIN J., KEOGH E., WEI L., LONARDI S.: Experiencing sax: a novel symbolic representation of time series. *Data Mining and Knowledge Discovery* 15, 2 (2007), 107–144. 1
- [LL18] LAFTCHIEV E., LIU Y.: Finding multidimensional patterns in multidimensional time series. In *KDD workshop on MiLeTS 2018, London* (2018). 1
- [LLS*20] LEE D. J. L., LEE J., SIDDIQUI T., KIM J., KARAHALIOS K., PARAMESWARAN A. G.: You can't always sketch what you want: Understanding sensemaking in visual query systems. *IEEE Trans. Vis. Comput. Graph.* 26, 1 (2020), 1267–1277. URL: <https://doi.org/10.1109/TVCG.2019.2934666>, doi:10.1109/TVCG.2019.2934666. 1, 2
- [LPH*20] LEKSCHAS F., PETERSON B., HAEHN D., MA E., GEHLENBORG N., PFISTER H.: Peax: Interactive visual pattern search in sequential data using unsupervised deep representation learning. In *Computer Graphics Forum* (2020), vol. 39(3), pp. 167–179. 1, 2
- [MA18] MANNINO M., ABOUZIED A.: Expressive time series querying with hand-drawn scale-free sketches. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (New York, NY, USA, 2018), CHI '18, Association for Computing Machinery, p. 1–13. URL: <https://doi.org/10.1145/3173574.3173962>, doi:10.1145/3173574.3173962. 2
- [NB05] NEGI T., BANSAL V.: Time series: Similarity search and its applications. In *Proceedings of the International Conference on Systemics, Cybernetics and Informatics: ICSCI-04, Hyderabad, India* (2005), pp. 528–533. 1
- [PKG11] PETITJEAN F., KETTERLIN A., GAŃCARSKI P.: A global averaging method for dynamic time warping, with applications to clustering. *Pattern Recognit.* 44, 3 (2011), 678–693. URL: <https://doi.org/10.1016/j.patcog.2010.09.013>, doi:10.1016/j.patcog.2010.09.013. 2
- [SLW*20] SIDDIQUI T., LUH P., WANG Z., KARAHALIOS K., PARAMESWARAN A. G.: Shapesearch: A flexible and efficient system for shape-based exploration of trendlines. In *Proceedings of the 2020 International Conference on Management of Data, SIGMOD Conference 2020, online conference [Portland, OR, USA], June 14-19, 2020* (2020), Maier D., Pottinger R., Doan A., Tan W., Alawini A., Ngo H. Q., (Eds.), ACM, pp. 51–65. URL: <https://doi.org/10.1145/3318464.3389722>, doi:10.1145/3318464.3389722. 2
- [TFRC07] TAK-CHUNG FU, FU-LAI CHUNG, ROBERT LUK, CHAKMAN NG: Stock time series pattern matching: Template-based vs. rule-based approaches. *Engineering Applications of Artificial Intelligence* 20, 3 (2007), 347–364. URL: <https://www.sciencedirect.com/science/article/pii/S0952197606001278>, doi:10.1016/j.engappai.2006.07.003. 1
- [vdBS*21] VAN DE SCHOOT R., DE BRUIN J., SCHRAM R., ZAHEDI P., DE BOER J., WEIJDEMA F., KRAMER B., HUIJTS M., HOOGWERF M., FERDINANDS G., HARKEMA A., WILLEMSSEN J., MA Y., FANG Q., HINDRIKS S., TUMMERS L., OBERSKI D. L.: An open source machine learning framework for efficient and transparent systematic reviews. *Nature Machine Intelligence* 3, 2 (2021), 125–133. doi:10.1038/s42256-020-00287-7. 2
- [YLC*19] YU C., LUO L., CHAN L. L.-H., RAKTHANMANON T., NUTANONG S.: A fast lsh-based similarity search method for multivariate time series. *Information Sciences* 476 (2019), 337–356. 2