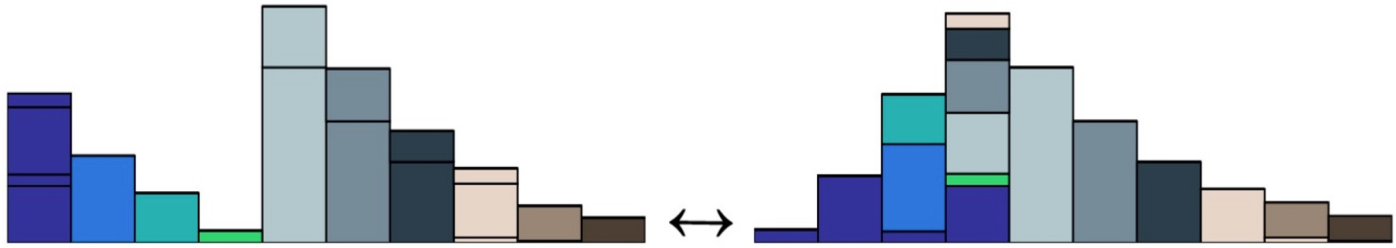
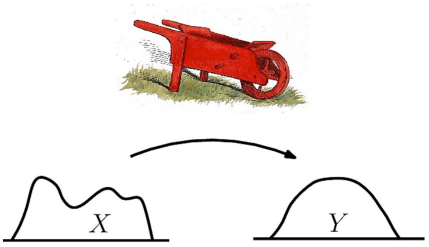
Master Project: Multivariate Time Series EMD Computation

University of Zurich[UZH]



Introduction

This project investigates efficient and scalable algorithms for computing the Earth Mover's Distance (EMD) on multivariate time series (MTS). EMD-based similarity analysis (EMDSA) is widely used in multimedia information retrieval and pattern recognition because it effectively measures similarity between complex distributions. However, applying EMD directly to high-dimensional multivariate time series is computationally expensive and memory intensive, making it difficult to scale to large datasets.



EMD computes the minimum cost of transforming one distribution into another, imagining moving dirt from one location to another.

The proposed research uses the public MagnaTagATune (MTAT) dataset, which contains approximately 5,000 audio tracks and more than 26,000 audio segments. Each segment is represented by approximately 250 acoustic features sampled over roughly 300 time steps, resulting in high-dimensional multivariate time series that require several gigabytes of storage. Computing pairwise similarities for tasks such as retrieval, clustering, and window-based analysis therefore becomes computationally prohibitive.

To address this challenge, the project investigates dimensionality reduction techniques that project the original feature space into lower-dimensional subspaces while preserving the essential similarity structure. Although dimensionality reduction significantly decreases storage requirements and computational cost, EMD computation remains expensive because its complexity increases rapidly with the number of histogram bins. Consequently, additional optimization strategies are required to achieve interactive or near real-time performance, for example, MapReduce [1], bulk synchronous parallel [2], [3] and resilient distributed datasets [4].

Beyond track-level similarity, the project also investigates segment-level similarity search. Instead of comparing complete audio tracks, the objective is to retrieve the most similar segments to a user-specified query segment across the entire database. Efficient segment retrieval requires algorithms capable of avoiding unnecessary EMD computations while maintaining high retrieval accuracy. To achieve this, the project will investigate boundary-based pruning techniques that eliminate unlikely candidate segments before performing full EMD evaluation.

The proposed methods will be evaluated in terms of computational efficiency and retrieval effectiveness. Particular emphasis will be placed on quantifying the trade-off between runtime performance and approximation accuracy, providing scalable EMD computation techniques for large-scale multivariate time series analysis.

Assignment

The project consists of four main components.

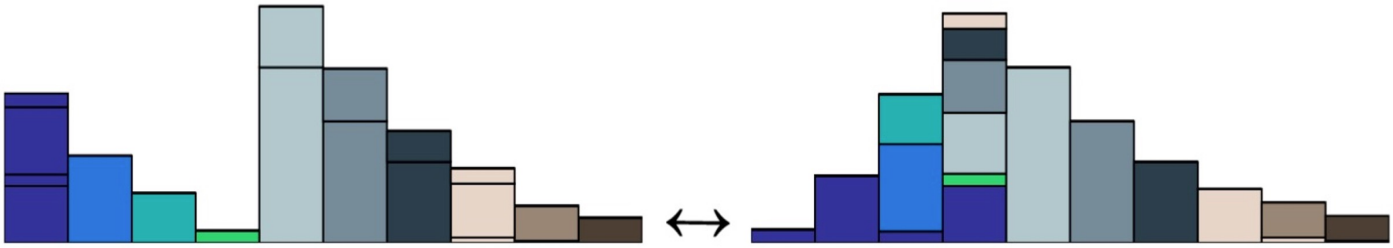
- Data Preprocessing and Dimensionality Reduction. The MagnaTagATune dataset needs an initial preprocessing pipeline reconstructs complete audio recordings by grouping fragments belonging to the same track, ordering them chronologically, and concatenating their corresponding feature matrices along the temporal dimension. To ensure balanced similarity computation, the extracted acoustic feature values are normalized using attribute-specific normalization schemes that preserve their statistical characteristics while preventing individual features from dominating the similarity measure. Principal Component Analysis (PCA) is then applied to project the normalized feature matrices into a lower-dimensional feature space. This reduces memory consumption and computational complexity while retaining the principal variation within the data.
- Track-level EMD Computation. The first similarity analysis operates on complete audio tracks. Pairwise similarities are computed using Earth Mover's Distance over the reduced feature representations. The objective is to investigate scalable EMD computation techniques that preserve similarity accuracy while substantially reducing computational cost.
- Segment-level EMD Computation. To capture localized musical characteristics, the project extends similarity analysis from complete tracks

VISUALIZATION AND MULTIMEDIA LAB

Professor Dr. Renato Pajarola
Visualization and Multimedia Lab
Department of Informatics
University of Zurich

Master Project: Multivariate Time Series EMD Computation

University of Zurich[UZH]



to individual time segments. Given a user-defined query segment, the system searches for the most similar segments across the entire database. Because exhaustive EMD computation for every candidate segment is computationally infeasible, efficient search strategies are investigated. In particular, boundary-based pruning techniques are used to estimate lower and upper similarity bounds, allowing unlikely candidate segments to be discarded before performing the full EMD computation. The effectiveness of different pruning strategies will be evaluated with respect to both computational efficiency and retrieval accuracy.

- Application. A web-based application will be developed to demonstrate and evaluate the proposed framework. The application will support interactive exploration of the processed dataset, including: Selection of specific audio tracks and individual segments; Segment-level similarity search and recommendation; Audio playback for qualitative comparison of retrieved results. The application will serve as a proof-of-concept platform integrating similarity computation, indexing, and interactive exploration into a unified system.

Project Type

Master Project

Remarks

The project description and requirements can be adjusted based on the size of the team. Additional tasks may be included, or the scope may be modified to focus on specific aspects of the project.

Requirements

Experience in programming with C++ and Python. Good knowledge of mathematics, linear algebra, and statistics.

Supervision

Prof. Dr. Renato Pajarola
Xiao Tan (Assistant)

Workload

- 40% Theory
- 40% Implementation
- 20% Evaluation

Contact

Email: xtan@ifi.uzh.ch (supervisor)
or timolennard.tietje@uzh.ch (team member)

References

[1] Kristen Grauman and Trevor Darrell. Fast contour matching using approximate earth mover's distance. In *Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2004. CVPR 2004.*, volume 1, pages I–I. IEEE, 2004.

[2] Jin Huang, Rui Zhang, Rajkumar Buyya, and Jian Chen. Melody-join: Efficient earth mover's distance similarity joins using mapreduce. In *2014 IEEE 30th International Conference on Data Engineering*, pages 808–819. IEEE, 2014.

[3] Jia Xu, Bin Lei, Yu Gu, Marianne Winslett, Ge Yu, and Zhenjie Zhang. Efficient similarity join based on earth mover's distance using mapreduce. *IEEE Transactions on Knowledge and Data Engineering*, 27(8):2148–2162, 2015.

[4] Jin Huang, Rui Zhang, Rajkumar Buyya, Jian Chen, and Yongwei Wu. Heads-join: Efficient earth mover's distance similarity joins on hadoop. *IEEE Transactions on Parallel and Distributed Systems*, 27(6):1660–1673, 2015.

VISUALIZATION AND MULTIMEDIA LAB

Professor Dr. Renato Pajarola
Visualization and Multimedia Lab
Department of Informatics
University of Zurich