Robust and Fast Similarity Search for Moving Object

Ritu Patel, Sarthak Siddhpura, Moin Vinchhi, Vrunda Patel, Vishv Boda
Ahmedabad University, Ahmedabad, Gujarat 380009, India

I. Research Contributions by Paper

- Introduced Edit Distance on Real sequence (EDR), a novel distance function for trajectory similarity, robust to noise and local time shifts by quantizing element-pair distances to 0 or 1 and minimizing edit operations.
- Demonstrated EDR's superior robustness over Euclidean distance, DTW, and ERP, and higher accuracy than LCSS on benchmark datasets (e.g., ASL, Slip, Kungfu) with noise and time shifting.
- Developed three pruning techniques—mean value Q-grams, near triangle inequality, and trajectory histograms—to enhance EDR retrieval efficiency without warping constraints, tested on datasets like NHL (5000 trajectories).
- Combined pruning methods (e.g., histogram pruning, Q-grams, triangle inequality) to reduce false candidates, achieving up to 20x speedup ratio over individual methods on large datasets (e.g., 100,000 random walk trajectories).
- Evaluated pruning efficiency via pruning power and speedup ratio, identifying optimal configurations (e.g., PS2 with Q-gram size 1, histogram bin size ϵ) across diverse trajectory lengths (30–1024).
- Validated EDR's efficacy using classification error rates on labeled datasets, showing improved nearest-neighbor prediction over existing measures under noisy conditions.

II. Key Insights From Paper

The paper by Chen et al. offers valuable insights for our research on identifying abnormal driving behavior using trajectory datasets and probabilistic-statistical methods, particularly through its introduction of the Edit Distance on Real sequence (EDR) metric. EDR's ability to handle noise and local time shifts—common in real-world GPS trajectories due to sensor errors or traffic variability—makes it a promising tool for defining "normal" behavior on road patches. Unlike Euclidean distance, which assumes point-to-point alignment and struggles with noise (e.g., high classification error rates in ASL data), EDR quantizes distances to 0 or 1 based on a threshold ϵ, effectively filtering outliers. For our binary classifier, this suggests a probabilistic approach where trajectories on a road patch are compared to a reference "normal" trajectory (e.g., a centroid from clustering). By computing EDR scores, we could establish a statistical distribution of distances, with normal behavior falling within a threshold (e.g., mean ± standard deviation) and abnormal behavior flagged as significant deviations—such as erratic lane changes or speed inconsistencies—reflected in higher edit operation counts. The paper's tests on benchmark datasets (e.g., Slip data) confirm EDR's robustness, reducing false positives in similarity searches, which aligns with our need for reliable anomaly detection in noisy naturalistic driving data.

A second key insight is the paper's pruning techniques, which enhance EDR's efficiency and inspire scalable anomaly detection. The mean value Q-grams method segments trajectories into subsequences, enabling fast similarity checks without predefined warping constraints, unlike DTW or LCSS. For our project, this could translate into a spatial-temporal feature extraction strategy: segmenting trajectories by road patch (e.g., 100-meter segments) and computing Q-gram-based similarity scores to a normal profile. The histogram pruning approach, particularly with one-dimensional sequences (pruning power up to 0.8 on Kungfu data), offers a statistical lens by binning trajectory features (e.g., x-y coordinates) and comparing distributions. We could adapt this by constructing histograms of normal behavior per patch—derived from a training set—and using a probabilistic divergence metric (e.g., KL divergence) to classify trajectories as abnormal if their histograms deviate significantly. The paper's finding that combining pruning methods (e.g., histograms followed by Q-grams) boosts speedup ratios (e.g., 5x over Q-grams alone on NHL data) suggests a hybrid approach for our classifier: initial histogram-based filtering followed by EDR-based refinement, ensuring computational feasibility for large-scale trajectory datasets while preserving detection accuracy.

Finally, the paper's emphasis on robustness and efficiency supports our goal of a practical, interpretable classifier. EDR's design penalizes unmatched trajectory segments, capturing spatial-temporal anomalies (e.g., sudden detours) that Euclidean or DTW might miss due to rigid alignment. This aligns with our assumption that drivers behave similarly on a road patch, allowing us to model normalcy as a low EDR score relative to a representative trajectory. The classification error rate evaluation (e.g., lower misses than LCSS on noisy data) validates EDR's potential as a similarity backbone for supervised learning, where labeled anomalies (e.g., from crash reports) train the classifier. The pruning techniques' independence—demonstrated by consistent pruning power across combination orders—offers flexibility to tailor our system to specific road patches or datasets. By integrating EDR with probabilistic thresholds (e.g., p-values from a distance distribution) and pruning-inspired feature reduction, we can develop a binary classifier that efficiently identifies abnormal driving behaviors—such as aggressive maneuvers or route deviations—enhancing road safety through precise, data-driven anomaly detection.

III. Unaddressed Issues and Assumptions

- **Unaddressed Issues:**
  - Lacks real-time implementation details, limiting applicability for dynamic anomaly detection in driving scenarios.
  - Does not address contextual factors (e.g., traffic, road conditions) that could redefine normal behavior, reducing classifier generalizability.
  - No discussion on handling multi-vehicle interactions within trajectories, critical for urban road patch analysis.
  - Evaluation focuses on similarity retrieval, not anomaly detection, leaving uncertainty in EDR's performance for binary classification tasks.
- **Assumptions Made:**
  - Assumes trajectories are two-dimensional and time-discrete, potentially overlooking speed or acceleration features vital for driving behavior.
  - Assumes noise and time shifts are the primary distortions, ignoring systematic biases (e.g., GPS drift) that could skew EDR scores.
  - Assumes uniform ϵ threshold suits all datasets, risking over- or under-sensitivity to anomalies across diverse road patches.
  - Assumes pruning methods preserve all relevant anomalies, possibly discarding subtle deviations in high-dimensional data.

IV. Motivation Behind Choosing This Paper

This paper was selected for its innovative EDR distance function and efficient pruning techniques, which directly support our research on abnormal driving behavior detection using trajectory datasets. EDR's robustness to noise and time shifts addresses real-world trajectory challenges, providing a reliable similarity measure to define normal behavior statistically on road patches. The pruning methods—Q-grams, histograms, and triangle inequality—offer scalable feature extraction and comparison strategies, crucial for processing large datasets efficiently. The paper's probabilistic underpinnings and experimental validation on diverse datasets (e.g., NHL, random walk) align with our prob-stat approach, inspiring a classifier that leverages EDR scores and statistical thresholds to distinguish normal from abnormal trajectories, enhancing safety through precise spatial-temporal analysis.

References