# Topological Trajectory Clustering with Relative Persistent Homology

Florian T. Pokorny, Ken Goldberg and Danica Kragic

*Abstract*— **Clustering techniques provide a key methodology to reason about databases of trajectories. In robotics, and within the learning from demonstrations framework in particular, robots require the ability to partition a set of demonstrations into meaningful subsets before learning motion primitives based on the identified classes. These trajectories typically lie in subsets of the environment, or within robot configuration spaces, that exhibit interesting topological properties. Current distance-based clustering techniques are however unable to take advantage of knowledge of the topology of the space within which the trajectories lie, making it difficult to encode hard constraints posed by such information. We propose an approach to trajectory clustering which is based on relative persistent homology and which can incorporate global constraints formalized in terms of the topology of sublevelsets of an arbitrary function. Our approach can be applied to sublevel sets of a probability distribution, allowing us to simultaneously reason about trajectories using probabilistic and topological information. Our experiments comparing the proposed method to distance based hierarchical clustering by a standard Fréchet distance indicate that topological clustering can provide a highly scalable alternative to metric clustering, enabling us to cluster more than 11.000 GPS trajectories with more than 5 million GPS points in less than 30s and with smaller memory requirements compared to a hierarchical Frèchet distance based clustering which took 57.7 minutes to determine clusters.**

## I. Introduction

The problem of reasoning about large databases of trajectory data obtained in domains such as robotics, computer vision, computer animation and from surveillance systems remains a key challenge in machine learning and robotics. In recent years, large trajectory datasets of GPS traces [25] as well as motion capture [9] have become available. Of particular interest to robotics is the emergence of cloud-based large-scale data sets of human and robot motion patterns that robots could access online within the paradigm of Cloud Robotics [15].

Some of the key difficulties in applying machine learning methods to large databases of trajectories arise from the fact that trajectories are naturally of varying length and hence not easily representable in a vector space of fixed dimension – which is the natural domain for popular methods such as Support Vector Machines. Furthermore, probabilistic approaches to motion analysis such as Gaussian Processes [19] are smooth in nature, often making the formulation of discrete constraints, such as whether a trajectory passes an obstacle to the right or to the left, difficult to encode. Topological methods hold promise in this respect because they are able to extract such discrete information about the global shape of data. Following the discovery of persistent homology [11], the incorporation of topological techniques with Machine Learning is now beginning to receive increased attention, for example at recent workshops at the flagship machine learning conferences ICML and NIPS [17], [24]. While trajectories can be clustered geometrically using several distance measures such as the Hausdorff or Fréchet metric, or using techniques such as string kernels [6], [30] and Dynamic Time Warping, the complexity of distance based approaches scale at least quadratically in the size of the trajectory dataset, making these methods challenging to apply with very large databases of trajectories. Furthermore, distance based clustering is rather sensitive to the chosen distance measure, with each such measure having benefits and drawbacks. The Hausdorff distance, for example, is highly sensitive to point-wise differences between trajectories, while $L_2$ based averaging approaches are on the other hand very insensitive to outliers, leading to ongoing research into trajectory distance measures [32].

In this work, we propose a *topological* rather than purely *geometric* approach to clustering trajectories with applications to robotics. We show in experiments with GPS trajectories that the proposed method can yield fundamentally different clusters compared to a clustering by Frèchet distance. We furthermore show how our approach can be combined with probabilistic reasoning by considering sublevel sets of a probability density for a pedestrian motion dataset.

## II. Background and Related Work

The processing, classification and clustering of trajectory data is a sizable research area. Some of the current methods are reviewed in [32]. In the robotics domain, trajectories, recorded for example as sequences of joint-angles of a robotic arm, play an important role in the *learning from demonstration* framework [1], [5], [26]. There, trajectory sequences are recorded during a demonstration phase, where a human instructor 'teaches' a robot how to perform a certain task. The robot then uses the trajectory data to model *motion primitives* that can be adapted based on environment conditions. In [13], the automated extraction of clusters of trajectories in order to obtain vocabularies of motion was studied using a system comprising filtering, segmentation and clustering using K-means. Jenkins [18], focussed on the extraction of behaviours for humanoid motion in particular.

The work of Knepper [20] studied classes of local path segments in order for a robot to reason about motion alter-

Florian T. Pokorny and Ken Goldberg are with the Department of Computer Science and Electrical Engineering, University of California, Berkeley. Ken Goldberg is also with the Department of Industrial Engineering and Operations Research, University of California, Berkeley. Danica Kragic is with CAS/CVAP, KTH Royal Institute of Technology ftpokorny@berkeley.edu, dani@kth.se, goldberg@berkeley.edu

Preprint submitted to 2016 IEEE International Conference on Robotics and Automation. Received September 15, 2015.

natives and is also related to the overall problem of trajectory clustering. Trajectories have been clustered using local sub-trajectories in the work of Lee [21] and also Buchin [7] who focused on extracting commuting patterns using the Fréchet distance which we will also use for comparison purposes.

Generally, previous clustering methods are either geometric in nature such as [21] or are based on probabilistic models, as in [14] who applied a mixture model for trajectory clustering and [23] who applied a hidden Markov model to detect activities from trajectories. Our work is distinct from these approaches, applying instead the topological techniques of persistent homology [11] to arrive at a trajectory classification which relies on *global topological information* extracted either from the trajectory data itself, or using a simplicial model of the environment containing these trajectories. Our current work extends our efforts [28] that introduced the use of persistent homology as a trajectory clustering technique, and showing how filtrations of Delaunay-Čech complexes can be utilized to cluster trajectories with fixed common start and end points. This work was recently extended to a construction in [29] to accommodate trajectories with varying start and end-points. This construction however relied on an enlarged simplicial complex construction, substantially increasing the size of the required simplicial data structure. Our present work generalizes [28] instead by means of *relative* persistent homology, allowing us to cluster trajectories with varying end-points based on the same simplicial complexes used for the classification of trajectories with fixed start and end points. The present work studies large datasets of real world GPS traces as a particular source of trajectory data which, as well as trajectories extracted from video [31], forms a common real data-source studied for example in anomalous event detection problems [27].

### A. Mathematical Background

The key tool used in this work is the machinery of *relative persistent homology* with coefficients in a field $\mathbb{F}$ which has emerged in the context of *persistence* [11] in recent years. This requires a few definitions from Algebraic Topology and is unfortunately rather technical in nature, requiring mathematical techniques which are difficult to fully explain and do justice in a short paper. The book [12] provides an excellent more complete introduction – we shall focus only on introducing the necessary key notation and ideas here.

We shall focus on persistence with binary coefficient field $\mathbb{Z}_2 = \{0, 1\}$, but more generally fields such as $\mathbb{Z}_p$ for some prime $p$ or the field of rational numbers $\mathbb{Q}$ could be considered. The binary field $\mathbb{Z}_2$ has the particular advantage of being simple and efficiently implementable on a computer. In 2 dimensions, the field $\mathbb{Z}_p$ allows us to distinguish, *e.g.* up to $p-1$-fold winding of trajectories around obstacles – $\mathbb{Z}_2$ in particular only enables us to detect whether we move to the left or right of voids/obstacles. The interested reader can consult [16], [12] for a detailed introduction to the required background material. We will focus on reviewing the main components of the approach.

*1) Simplicial Complexes:* We review the notion of a simplicial complex which represents the key data structure used in this work. A geometric $k$-simplex $\sigma = [v_0, \ldots, v_k]$ in $\mathbb{R}^d$ is a convex hull of $k+1$ affinely independent ordered points $v_0, \ldots, v_k \in \mathbb{R}^d$. We call $k$ the dimension of a $k$-simplex. If $\tau \subseteq \sigma$, $\tau, \sigma \in \mathcal{K}$, then $\tau$ is called a face of $\sigma$. In the special case of $\mathbb{Z}_2$ homology, the ordering can in fact be ignored. A geometric simplicial complex $\mathcal{K}$ is a non-empty set of simplices such that if $\sigma \in \mathcal{K}$ and $\emptyset \neq \tau \subseteq \sigma \in \mathcal{K}$, then $\tau \in \mathcal{K}$ and the intersection of any two simplices $\sigma, \tau \in \mathcal{K}$ is a face of both $\sigma$ and $\tau$. We write $|\mathcal{K}|$ for set of points in $\mathbb{R}^d$ contained in the union of all simplices in $\mathcal{K}$. The set $|\mathcal{K}|$ is a topological space with the subspace topology from $\mathbb{R}^d$. A subset of simplices $\mathcal{A} \subset \mathcal{K}$ that is itself a simplicial complex is called a subcomplex of $\mathcal{K}$. Note that 0-simplices just correspond to points, 1-simplices are finite line segments and 2-simplices are triangles in $\mathbb{R}^d$. The conditions in the definition of a simplicial complex ensure that the simplices are assembled in a natural manner and the notion of a simplicial complex generalises both the notion of a geometric graph and a triangulation.

*2) Relative Homology:* For a field $\mathbb{F}$, a $p$-chain $c$ is a formal sum $c = \sum_{i=1}^{k} \lambda_i \sigma_i$ of $p$-simplices $\{\sigma_i\}_{i=1}^{k} \subset \mathcal{K}$ with $\lambda_i \in \mathbb{F}$ and $C_p(\mathcal{K})$ denotes the $\mathbb{F}$-vector space of all $p$-chains. In particular, for finite simplicial complexes, 1-chains are finite linear combinations of edges and 2-chains are finite linear combinations of triangles. When no confusion arises, we write $C_p$ for $C_p(\mathcal{K})$ to simplify notation. For every geometric $p$-simplex $\sigma = [v_0, \ldots, v_p]$ let $\partial_p \sigma$ be the $p-1$-chain $\partial_p \sigma = \sum_{i=0}^{n} (-1)^i [v_0, \ldots, v_{i-1}, v_{i+1}, \ldots, v_p]$ consisting of a signed sum of faces of $\sigma$. For each $p \in \{0, \ldots, d\}$, $\partial_p$ extends to a linear map $\partial : C_p \to C_{p-1}$, called the boundary operator.

A $p$-chain $c$ such that $c = \partial_{p+1} \omega$ for some $\omega \in C_{p+1}$ is called a $p$-boundary. And a $p$-chain $c$ such that $\partial_p c = 0$ is called a $p$-cycle. The vector spaces of $p$-boundaries and $p$-cycles are denoted $B_p$ and $Z_p$ respectively. For a 1-chain $c$ corresponding to an oriented path from $s$ to $t$, $\partial c = t - s$. We have $\partial c = 0$, so that $c$ is a 1-cycle, for any closed oriented 1-chain $c$. Similarly, the boundary $\partial w$ of a 2-chain $w$ corresponding to an oriented collection of triangles (2-simplices) exactly corresponds to the oriented geometric boundary of these triangles (modulo $\mathbb{F}$) - the name *boundary operator* is hence also geometrically natural.

The $p$-th homology group of $\mathcal{K}$ is defined by $H_p(\mathcal{K}) = Z_p / B_p$. For a cycle $c \in Z_p$,we denote by $[c] \in H_p$ the resulting element in homology. Note that each $p$-cycle $c$ yields an element in $H_p$, but this representative is only unique up to elements in $B_p$. We are interested in $H_1$ in particular, which consists of equivalence classes of closed 1-cycles up 1-cycles that are boundaries of 2-cycles. In the case of $\mathbb{F} = \mathbb{Z}_2$, we can visualize 1-chains as a collection of edges in $\mathcal{K}$ which have non-zero coefficients in the chain. See Fig. 1 for an example.

The importance of homology in mathematics arises from the fact that it captures *global topological properties* about the topological space defined by $|\mathcal{K}|$. In particular homology

Preprint submitted to 2016 IEEE International Conference on Robotics and Automation. Received September 15, 2015.

remains invariant under continuous deformations of the space $|\mathcal{K}|$ (homotopies of $|\mathcal{K}|$). In particular, $b_p = \dim(H_p(\mathcal{K}))$ is called the $p^{th}$ Betti number and counts the number of connected components ($b_0$), tunnels ($b_1$), and higher dimensional voids in $|\mathcal{K}|$. The left part of Fig. 1 illustrates an example 1-cycle $c$ lying in a simplicial complex $\mathcal{K}$ and forming a basis of $H_1(\mathcal{K})$ which is in this case 1-dimensional and where we pick $\mathbb{F} = \mathbb{Z}_2$ coefficients. There, $|\mathcal{K}|$ is in fact homotopy equivalent to a circle and $dim(H_1(\mathcal{K})) = 1$.

In our prior work [28], we used a basis for $H_1(\mathcal{K})$ to topologically cluster trajectories $\alpha_0, \ldots, \alpha_n$, represented as edgepaths in $\mathcal{K}$. There, it was initially necessary to assume that all trajectories had the same fixed start and end vertices $s, t \in \mathcal{K}$ respectively in order to form $\mathbb{Z}_2$-cycles $c_i = \alpha_0 + \alpha_i$ that could then be classified in homology. We were only able to consider more general trajectories with general initial and terminal regions $S, T \subseteq \mathcal{K}$, by introducing a cone construction [29] which significantly increased the size and complexity of the approach.

We now review *relative homology*, which provides an alternative to standard homology: For a subcomplex $\mathcal{A} \subset \mathcal{K}$, we define the quotient vector space of relative $p$-cycles $C_p(\mathcal{K}, \mathcal{A}) = C_p(\mathcal{K})/C_p(\mathcal{A})$ The boundary operator descends to a linear operator on relative chains. We denote it by $\hat{\partial}_p$ [16]. We define the set of relative $p$-boundaries by $B_p(\mathcal{K}, \mathcal{A}) = im(\hat{\partial}_{p+1} : \mathcal{C}_{p+1}(\mathcal{K}, \mathcal{A}) \to \mathcal{C}_p(\mathcal{K}, \mathcal{A}))$. These correspond to $p$-chains $c \in C_p(\mathcal{K})$ such that $c = \partial_{p+1} w + a$ for some $a \in C_p(\mathcal{A})$ and $w \in C_{p+1}(\mathcal{K})$. Similarly, relative $p$-cycles are defined by $B_p(\mathcal{K}, \mathcal{A}) = ker(\hat{\partial}_p : \mathcal{C}_p(\mathcal{K}, \mathcal{A}) \to \mathcal{C}_{p-1}(\mathcal{K}, \mathcal{A}))$ and correspond to $p$-chains whose boundaries lie in $C_{p-1}(\mathcal{A})$. In particular, a relative 1-cycle $c$ is a sequence of edges such that start and end-point lie in $C_1(\mathcal{A})$. For a 1-chain corresponding to an oriented edge-path in $\mathcal{A}$, this occurs either when the path is cyclic (zero boundary), or when its initial and terminal points lie in $\mathcal{A}$. Finally, we have $B_p(\mathcal{K}, \mathcal{A}) \subseteq Z_p(\mathcal{K}, \mathcal{A})$ and the $p$-th relative homology is defined by $H_p(\mathcal{K}, \mathcal{A}) = Z_p(\mathcal{K}, \mathcal{A})/B_p(\mathcal{K}, \mathcal{A})$, describing equivalence classes of relative $p$-cycles modulo relative $p$-boundaries.

Relative homology is of importance since it allows us to consider properties of topological quotient spaces. Consider a simplicial complex $\mathcal{K}$ and a subcomplex $\mathcal{A} \subset \mathcal{K}$. We can consider identifying all of $\mathcal{A}$ with a single point – precicely, this corresponds to the topological quotient space $\mathcal{K}/\mathcal{A}$, within which a point $[x]$ corresponds to an equivalence class of a point $x \in \mathcal{K}$ modulo $\mathcal{A}$ and all of $\mathcal{A} \subset \mathcal{K}$ is identified with a single point in the quotient. An important result [16] states in particular that the first reduced homology of $\mathcal{K}/\mathcal{A}$ with field coefficients can be computed using the relative homology $H_1(\mathcal{K}, \mathcal{A})$, so that we can think of relative first homology in terms of the first homology of the quotient space. The two parts of Fig. 1 illustrate a basis for $H_1(\mathcal{K})$ and $H_1(\mathcal{K}, \mathcal{A})$, where $\mathcal{A}$ is the subcomplex of $\mathcal{K}$ containing all simplices of $\mathcal{K}$ whose vertices all lie in the shaded region (which is $|\mathcal{A}|$). Here, $\dim(H_1(\mathcal{K})) = 1$, while $dim(H_1(\mathcal{K}, \mathcal{A})) = 2$.

*3) Filtrations:* Simplicial complexes have for many decades enjoyed popularity in pure mathematics [16] in order to model and approximate various spaces of interest. Only recently however has there been interest in constructing simplicial complexes from real world data which has lead to the development of *persistent homology* [11], [8] which studies the homology of an increasing sequence (a filtration) of topological spaces. Persistence shares its origins with Morse theory, where one studies the topology of sublevel (or superlevel) sets of a function $f : X \to \mathbb{R}$ defined on a topological space $X$. Each sublevel set $X_r = f^{-1}((-\infty, r])$ yields a topological space $X_r$, where $X_r \subseteq X_{r'}$ whenever $r \leqslant r'$.

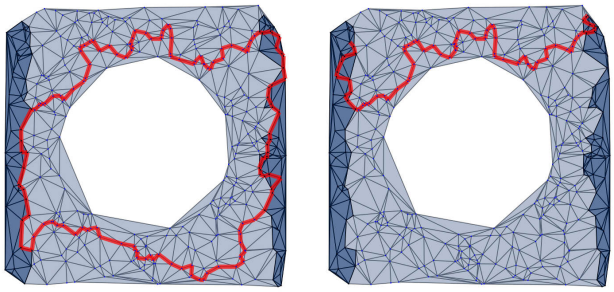

Fig. 1. We display a simplicial complex $\mathcal{K}$ in light and dark blue and a subcomplex $\mathcal{A} \subset \mathcal{K}$ in dark blue. The closed red edge-path $c_1$ on the left is a 1-cycle which cannot be represented as the boundary of a 2-cycle due to the hole in $\mathcal{K}$. In fact $[c_1] \in H_1(\mathcal{K})$ forms a basis for the 1-dimensional first homology group $H_1(\mathcal{K})$. On the right, we display a relative 1-cycle $c_2$ in red. Both $c_1, c_2$ in fact form relative 1-cycles and $[c_1], [c_2] \in H_1(\mathcal{K}, \mathcal{A})$ yield a basis for the 2-dimensional first relative homology group $H_1(\mathcal{K}, \mathcal{A})$. While $|\mathcal{K}|$ is homotopy equivalent (deformable) to a circle, the quotient space $\mathcal{K}/\mathcal{A}$ is homotopy equivalent to two circles glued at a common point, called a wedge of two circles - for this imagine first gluing the two shaded connected components of $\mathcal{A}$ together and then shrinking the resulting cylinder with a cut out hole until we obtain two circles. Since a wedge of 2 circles has a 2-dimensional $H_1(\mathcal{K}/\mathcal{A})$, we can also reason geometrically to understand why $dim(H_1(\mathcal{K}, \mathcal{A})) = dim(H_1(\mathcal{K}/\mathcal{A})) = 2$.

As $r$ increases, homological features can be 'born' and disappear or 'die' as the threshold $r$ increases. Persistence provides a computational mechanism for understanding these changes. To make this precise, we work with a filtration $\mathbb{K}$ of finite simplicial complexes in $\mathbb{R}^d$, by which we mean a sequence $\mathbb{K} : K_1 \subset K_2 \subset \ldots \subset K_n = K_\infty$ of finite simplicial complexes. Typically, each filtration index $i$ is associated to a real valued filtration value $r$ so that $K_i = f^{-1}(-\infty, r]$. For example, we can assign an arbitrary real valued scalar to each vertex of $K_n$, then the function $f(\sigma) = \max_{i \in \{0,\ldots,k\}} f(v_i)$ for an arbitrary $k$-simplex $\sigma = [v_0, \ldots, v_k] \in K_n$ yields a valid filtration function giving rise to a filtration of simplicial complexes when its sublevel sets are considered. When $K_i = f^{-1}((-\infty, r])$ we call $r$ the filtration value associated to the filtration index $i$.

### B. Simplicial complexes from data

To study homological features of a point-cloud dataset $X = \{x_1, \ldots, x_n\} \subset \mathbb{R}^d$, we need to construct a sequence of topological spaces modeling $X$. We will in particular consider the family of union of balls spaces $X_r = \bigcup_{x \in X} \{y \in \mathbb{R}^d : \|x - y\| \leqslant r\}$, for $r \geqslant 0$. For each $r$, $X_r$ is homotopy equivalent to the Delaunay-Čech complex $DC_r(X)$ [3], which is a simplicial complex defined for any

Preprint submitted to 2016 IEEE International Conference on Robotics and Automation. Received September 15, 2015.

finite set $X \subset \mathbb{R}^d$ where each subset of $d+1$ point is affinely independent. The assumption of affine independence is generic in that a uniform random sample satisfies this condition with probability one and we can also enforce the condition by an arbitrarily small perturbation of $X$. Let $D(X)$ denote the simplicial complex corresponding to the Delaunay triangulation of $X$ with simplices defined by $D(X) = \{[v_0, \ldots, v_k] : v_i \in X, \cap_{i=0}^{k} V_{v_i} \neq \emptyset\}$, where $V_x$ denotes the Voronoi cell containing $x$. For each $k$-simplex $\sigma = [v_0, \ldots, v_k] \in DC(X)$, define $f(\sigma) = \min\{r : \bigcap_{i=1}^{k} \mathbb{B}_r(v_i) \neq 0\}$, where $\mathbb{B}_r(x) = \{y \in \mathbb{R}^d : \|x - y\| \leqslant r\}$. The Delaunay-Čech complex $DC_r(X)$, for $r \geqslant 0$ is the sub-complex of $D(X)$ defined by $DC_r(X) = f^{-1}((-\infty, r])$. Since $DC_r(X)$ is homotopy equivalent to $X_r$, we can compute topological information about $X_r$ from $DC_r(X)$ at all scales $r \geqslant 0$. If a function $f : X \to \mathbb{R}$ is defined on the data $X$ – for example a log likelihood, probability density or cost function, we can furthermore define an induced filtration by extending $f$ to a function $f(\sigma) = \max_{i \in \{0,\ldots,k\}} f(v_i)$ for the $k$-simplex $\sigma = [v_0, \ldots, v_k]$ of $D(X)$, yielding a filtration $F(f)_r$ of simplicial complexes by the sub-levelsets of $f$. As we increase the threshold parameter $r$ a larger and larger sub-complex of $D(X)$ is considered. Super-levelsets can also be studied by replacing $f$ with $-f$. Fig. 2 illustrates examples of $DC_r(X)$ at various thresholds. Besides $DC(X)$, there furthermore exist alternate constructions yielding simplicial complexes, such as the Vietoris-Rips and Witness complexes [8] which are scalable to high-dimensional spaces, but which are not necessarily homotopy equivalent to $X_r$ at filtration level $r$.

### C. Persistent relative homology

When we apply homology to a filtration of simplicial complexes $K_1 \subset \ldots \subset K_n$, we obtain a sequence of linear maps $f_p^{i,j} : H_p(K_i) \to H_p(K_j)$ for $i \leqslant j$ induced by the inclusions $K_i \subset K_j$. The $p$-th persistent homology group for $i \leqslant j$ is given by $H_p^{i,j} = im f_p^{i,j}$, so that non-trivial elements in $H_p^{i,j}$ correspond to homology classes born at or before index $i$ and which survive until at least index $j$. The difference $j-i$ is called the *index* persistence of such a class. For us, $K_i = f^{-1}((-\infty, r_i])$, and $r_j - r_i$ is the persistence of the class. In fact, all the persistent homology groups can be computed by a decomposition of the persistence module into interval modules [10]. The $p$-th persistence diagram, captures the information about the birth and death of $p$-th homology classes as the filtration value increases. It consists of multisets of points in the extended upper left quadrant. Each point $(r_i, r_j)$ in the diagram corresponds to a homology class born at index $i$ and surviving until index $j$. Points that lie far above the diagonal have a large persistence and are hence considered important features distinct from smaller scale features due to noise. An example is presented in Fig. 2. Classes born at index $i$ and which do not die at the final filtration index $n$ are called essential and are associated to points of the form $(r_i, \infty)$ in the plane, extended formally to $(\mathbb{R} \cup \{\infty\})^2$. Note that the dimension of $H_1(K_i)$ for any $r_i > 0$ is equal to the number of points above and to the left

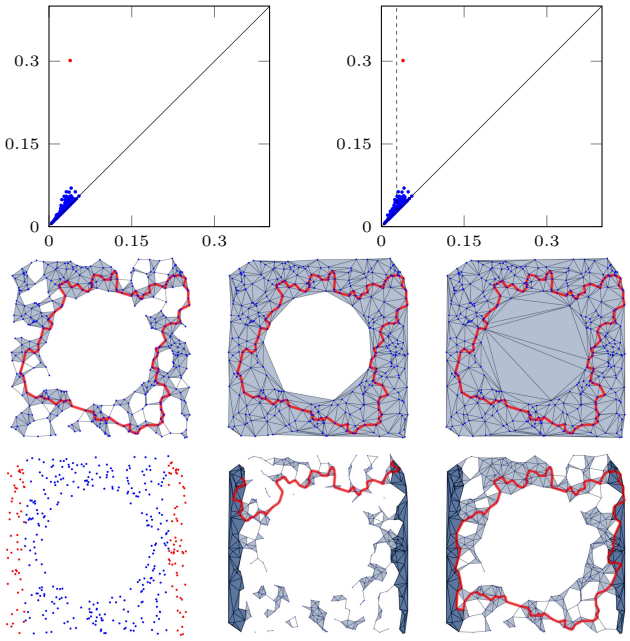


Fig. 2. Given the pointcloud $X \subset [0,1]^2$ in the bottom left, we consider $DC(X)$ whose first persistence diagram is displayed in the top left. A large persistence interval $(0.04, 0.30)$ is marked in red and corresponds to the hole in the point-cloud. The middle row displays $DC_r(X)$ at $r = 0.04$ when the large hole is first enclosed by simplices, at $r = 0.15$ and at $r = 0.30$ when the hole finally is covered by simplices. The red curve is a 1-cycle $c$ such that $[c]$ corresponds to this interval. The bottom middle and right figure displays $DC_r(X, Y)$, where $Y \subset X$ is the set of points marked in red in the bottom left and for $r = 0.03$ (middle) and $r = 0.04$. The top right figure shows the first persistence diagram of $DC_r(X, Y)$ relative to the shaded dark blue subcomplex $A(X, Y)$. The diagram differs only in minor detail from the persistence diagram of $DC(X)$ on the left - apart from the existence of a non-finite point $(0.03, \infty)$ indicated by the dashed line. For $r \in (0.04, 0.3)$, $dim(H_1(DC_r(X, Y), A(X, Y))) = 2$, as discussed in Fig. 1.

of $(r_i, r_i)$. To compute a basis for the persistent homology groups, we first assume without loss of generality that the filtration $\mathbb{K}$ has been refined to a simplex-wise filtration, where $K_j = \bigcup_{i=1}^{j} \sigma_i$, so that $K_{j+1} = K_j \cup \{\sigma_{j+1}\}$ and we hence add a single simplex in each step of the filtration. Given such a simplex-wise filtration, several algorithms (see *e.g.* [4]) are available to compute a basis of the persistent homology groups. We shall use the library [4] and the left-to right reduction algorithm described in [10] for this purpose. Fig. 2 illustrates an example of a filtration and an associated homology basis.

Persistence has recently emerged as a new approach in data analysis since large persistence intervals in the persistence diagrams are provably stable under noise [11] and represent global structure information about the point-cloud which is not extractable with non-topological methods. In this work, we will use a less-common extension of persistence to relative homology. For a sub-complex $A \subseteq K_1$ of a filtration, we consider the sequence of linear maps on *relative* homology $\hat{f}_p^{i,j} : H_p(K_i, A) \to H_p(K_j, A)$ for $i \leqslant j$ induced by the inclusions $K_i \subset K_j$. A basis for these relative homology groups can be obtained by a modification of the standard left-to right reduction algorithm. Since $H_1(K_i, A) \simeq H_1(K_i/A)$, we can in particular think

Preprint submitted to 2016 IEEE International Conference on Robotics and Automation. Received September 15, 2015.

of the first relative persistent homology groups as measuring the persistence of the sequence of topological quotient spaces $K_i/A$ for $i \in \{0, \ldots, n\}$.

An approach we introduce in his work is to define $A(X, Y)$ to be the sub-complex of $DC_\infty(X) = D(X)$ of simplices whose vertices are contained in $Y$ (shaded in Fig. 2). We then augment $A(X, Y)$ to a filtration $DC_r(X, Y)$ by inserting the remaining simplices of $D(X)$, with filtration order defined by the defining function $f$ of $DC(X)$. In the top left of Fig. 2, we illustrate the resulting relative first persistence diagram for $H_1(DC_r(X, Y), A(X, Y))$.

## III. Methodology

As input, we consider a set $\Gamma = \{\gamma_1, \ldots, \gamma_m\}$ of piecewise linear trajectories $\gamma_i : [0, 1] \to \mathbb{R}^d$ which can have arisen for example as GPS traces, or as joint-configuration trajectories of a robotic system. We assume furthermore that these trajectories all start and end in some specified (possibly disconnected) region $R \subset \mathbb{R}^d$.

### A. Discretization and Setup

In order to apply our simplicial complex based approach, we require a simplicial complex $L$ such that each trajectory $\gamma_i$ can be discretized as a sequence of 1-simplices in $L$. For a simplicial complex such that $\Gamma \subset |L|$ (recall $|L|$ denotes the union of all simplices in $L$) this discretization can be performed by first mapping each vertex of the piecewise linear $\gamma$ to its nearest 0-simplex in $L$ and to then connect each consecutive pair of such mapped trajectory points by a shortest path of 1-simplices in $L$. Given $R$, define $A \subset L$ to consist of those simplices of $L$ whose vertices all lie in $R$ and we assume that the mapped initial and terminal positions of the trajectories also map to corresponding vertices in $A$ under our discretization procedure.

Given, $A, L$, suppose now that there exists a filtration $L = K_1 \subset K_2 \subset \ldots \subset K_n$ of simplicial complexes. In that case, we can consider the first persistent relative homology groups arising from the inclusions

$$H_1(K_1, A) \to \ldots \to H_1(K_n, A).$$

Each $\gamma \in \Gamma$ forms a relative cycle in $Z_1(K_i, A)$ *for all* $i \in \{1, \ldots, n\}$. For each fixed $i$, we can furthermore consider $[\gamma] \in H_1(K_i, A) = Z_1(K_i, A)/B(K_i, A)$. The key insight of persistence [11] however implies that we can compute a basis for all $H_1(K_i, A)$, $i = 1, \ldots, n$ *simultaneously*, and the algorithms of persistence [4] furthermore provide the most efficient known implementation of determining these bases *even when* $i$ *is fixed*. To compute these bases, we first create an arbitrary simplex-wise filtration of $L = K_1$ and augment this filtration by adding each of the simplices of $K_n$ such that simplices of $K_i$ are inserted before $K_j$ and such that the faces of any simplex $\sigma$ are inserted before $\sigma$. This results in a refined filtration $M_1 \subset \ldots \subset M_p = K_n$, where one simplex is inserted at each step and $A = M_s$ for some $s$. By applying the standard left-to-right matrix reduction to the boundary matrix $\partial$ of this filtration, one arrives at a reduction $R = \partial V$ such that the required bases for $H_1(K_i, A)$ consists of a subset of lower parts of columns of $R$ and $V$ below row $s$. The column indices are specified by the persistence indices (see also [28] for the non-relative version of this). For each fixed filtration index $i$, we thus obtain a basis $b_1, \ldots, b_w$ of $H_1(K_i, A)$ allowing us to compute $\mathbb{F}$-coordinates of the image of our trajectories in $H_1(K_i, A)$.

Note that this procedure amounts to a finite dimensional binary vectorial featurization of the trajectories. However, unlike other known featurizations, the dimensionality of the features is independent of the length of the trajectory and instead depends on the global topology of the quotient space $K_i/A$ modelled by the simplicial complex.

We propose to cluster trajectories with the same coordinates into a joint cluster for each filtration index $i$. As $i$ is increased these trajectory clusters then merge hierarchically until some final filtration value. The trajectories however do not necessarily all merge into a single cluster at the final filtration index (because $H_1(K_\infty, A)$ need not be trivial). Note that, compared to [28], relative homology has the additional benefit that each trajectory *already forms a relative cycle*, while in the cited work, the authors needed to apply concatenations of trajectories to form (non-relative) cycles in $H_1(K_i)$.

It follows from the fact that $H_1(K_i, A) \simeq H_1(K_i/A)$ that two trajectories $\gamma, \gamma'$ such that $[\gamma] \neq [\gamma'] \in H_1(K_i, A)$ are not continuously deformable to one another if we allow their their end-points to also vary continuously in $|A|$ – this a relaxed version of the standard notion of homotopy, where the end-points need to be fixed. Our clusters hence consist of sets of trajectories such that no trajectory from one cluster can be continuously deformed to any trajectory of the other cluster in this sense. The converse is however not necessarily true: it can happen that two trajectories in the same cluster cannot be continuously deformed to one another in $|K_i/A|$. Besides this information, the clustering is hierarchical in nature and we can extract from the persistence intervals the filtrations at which two trajectories remain in separate clusters and when they merge.

*1) Filtrations of interest:* The freedom in choice of filtration allows us a wide modeling capability to express *constraints*. Given just a point-cloud $X$ and terminal subset $Y \subset X$, a natural choice is the filtration given by $DC_r(X, Y)$ introduced earlier and shown in Fig. 2. Using the persistence intervals, we can automatically determine filtration values such that our trajectories lie in $|DC_R(X, Y)|$ for some minimal $R > 0$ and such that $DC_R(X, Y)$ is path connected (0-persistence). We can furthermore identify filtration regions where the remaining Betti numbers (number of voids, tunnels, etc.) do not vary under noise (*i.e.* large persistence intervals) using the persistence stability theory [11]. Two particular situations to distinguish are in particular the case where $X$ is a large and dense set of samples from some configuration space $\mathcal{C}$, in which case $DC_R(X)$ can be expected to be homotopy equivalent to $\mathcal{C}$ and the classification of trajectories depends *on the topology of* $\mathcal{C}$ *only*. In the second case, $X$ is sparse and might only consist of trajectory points itself – in that case the classification of

Preprint submitted to 2016 IEEE International Conference on Robotics and Automation. Received September 15, 2015.

trajectories captures *global intrinsic information about the shape of the union of these trajectories themselves* – recall here that $DC_r(X)$ is topologically the same (homotopy equivalent) to the union of ball space $X_r$ and hence recovers topological information of $X_r$ as these balls grow around the samples $X$. We will work with both cases, as has been done in the non-relative case in [28].

Note that in the above, only the filtration function $f$ defining $DC_R(X)$ was used. In general, however our filtration can arise as sublevel sets of an *arbitrary function*. In particular, we can consider a probability density $f$, log-likelihoods, cost functions, etc. Our classification approach enables us to encode topological constraints by considering the homological properties of all sublevelsets simultaneously. The notion of topology is hence rather generic in nature as a sublevelset of some probability density on a simple topological space, such as a square, can have intricate features, while the square itself is topologically trivial. Similarly $DC_R(X, Y)$, for sparse $X \subset [0, 1]^2$, can capture information about $X$ itself, rather than the square's topology.

### B. Classifying ship movements

## IV. Experiments

We work with a dataset of GPS ship trajectories [22] from a region around Shanghai, one of the world's busiest shipping regions, which was recorded during 6 months in 2013. We are interested in classifying the motion of ships between the shaded northern and southern region in Fig. 3. Our dataset consists of 11711 trajectories with 5102072 GPS data points which corresponded to ships traveling between these regions and such that there was no more than 30 minutes of time difference between each consecutive GPS signal point along a trajectory. We now demonstrate how our approach benefits from information about the geometry of the sea region. We hence assume the knowledge of a sea/land classification of this area is available and create a detailed map by uniformly sampling 100000 points $X \subset \mathbb{R}^2$ lying in the water region. Denoting the subset of $X$ lying in the shaded region by $Y$, we constructed a triangulation of the water region using $DC_r(X, Y)$ for $r = 0.0002$ longitude/latitude degree. Next we discretized the trajectories as approximate edge-paths in $DC_r(X, Y)$ as described in Sec. III. See Fig. 3 for an illustration. Using the filtration $DC_r(X, Y)$, we computed relative persistent homology coordinates for all trajectories to cluster the dataset. The filtration $DC(X, Y)$ was constructed in $1.15s$ and, once each trajectory was represented as a cycle, it took less than 0.002s per trajectory to determine its relative persistent homology coordinates. We obtained a total of 15 classes at $r = 0.0002$. All classes are displayed in the top of Fig. 4, and we can see that this clearly corresponds to a classification of the ship movements relative to the various islands. Note the differences with a metric classification: our classification allows for outliers in this setting as it depends only on topological properties of the water region. The middle row of Fig. 4 displays two example classes. In the bottom right, we see that a single linkage clustering by Fréchet distance does not result in clustering representing

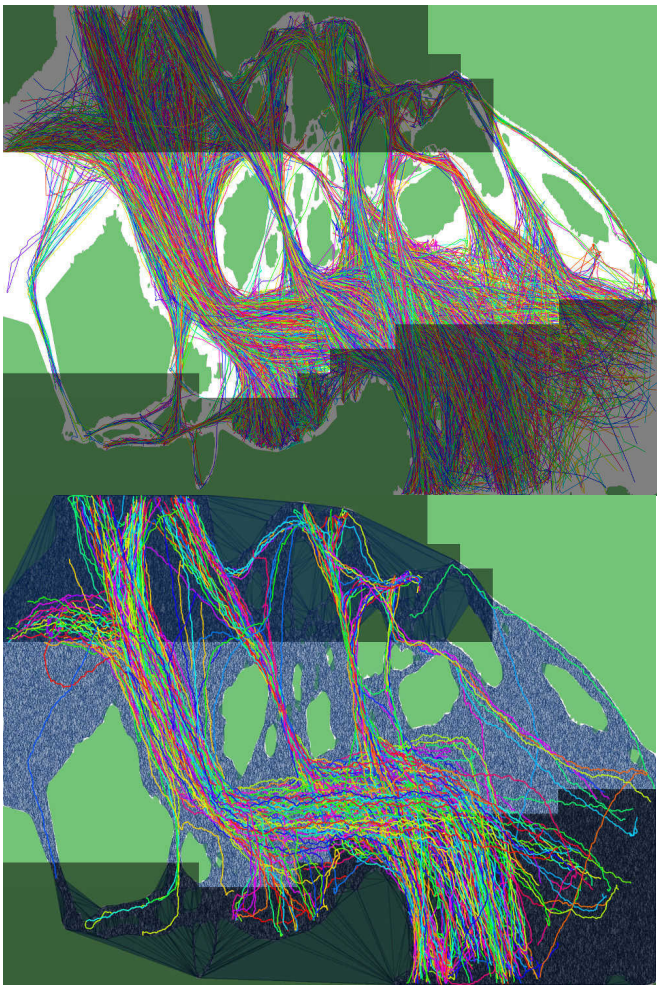

Fig. 3. A dataset of 11711 ship GPS traces of vessels traveling between the shaded regions around Shanghai is shown in the top figure. The bottom figure displays $DC_r(X, Y)$ and a subset of 500 trajectories discretized as relative 1-cycles in the simplicial complex representing the water area at $r = 0.0002$.

the 'hard constraints' enforced by the environment topology as these methods do not incorporate such information. The Fréchet clustering took 57.7 minutes and approached the $8GB$ memory limit of the laptop used. The topological clustering, for fixed sample size $X$, on the other hand had constant memory and time requirements in the number of trajectories, is easily parallelizable, and took only less than 30 seconds, allowing us to scale our approach to large numbers of trajectories.

### A. Application to traffic anomaly detection

We now focus on a traffic anomaly detection/classification problem, where we assume that we are not given any road/land classification data but only the trajectories themselves. We consider the set of 1685 GPS traces of cars driving across a highway crossing next to Frankfurt airport, obtained from OpenStreetMap [25]. The dataset contains 60987 GPS points $X$ and we furthermore select the subset of 177 trajectories that start and end in the shaded regions

Preprint submitted to 2016 IEEE International Conference on Robotics and Automation. Received September 15, 2015.

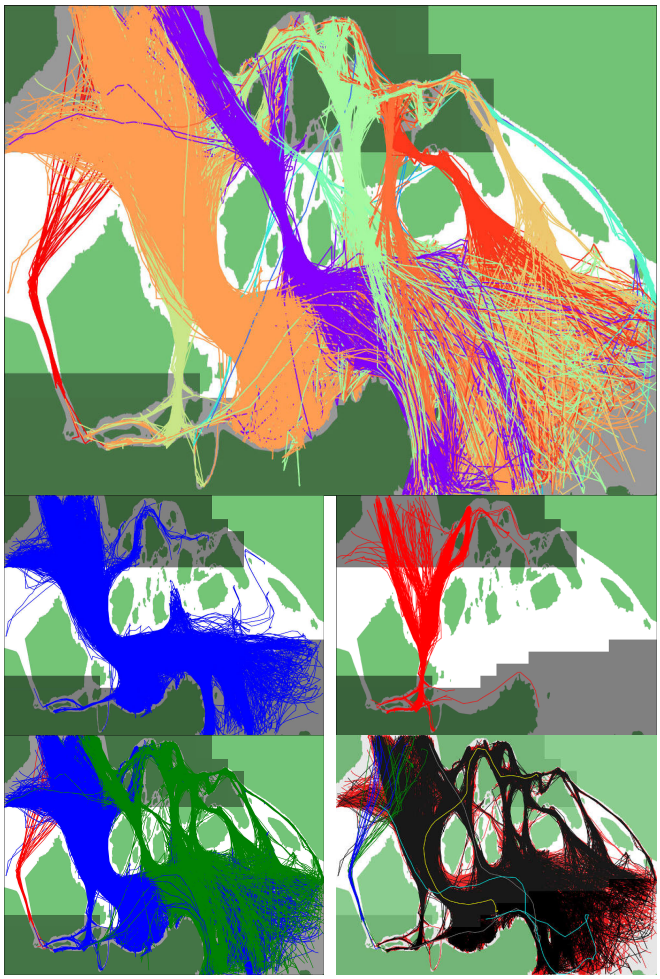

Fig. 4. The top figure displays all 15 found clusters at $r = 0.0002$, while the middle row illustrates two classes in isolation. Note how much variation these classes exhibit - yet they are distinguishable by the 'hard environment constraint' posed by the small southern island that they pass in a distinct manner. The bottom right displays a single linkage clustering by discrete Fréchet distance at distance 0.055, yielding 7 completely environment agnostic clusters. The bottom right figure shows the classification using our method at a higher filtration value of $r = 0.01$, where the smaller islands have been covered by simplices, resulting in only 3 trajectory classes at that filtration level.

shown in Fig. 5. Using only $X$, and the subset $Y$ of points in the shaded regions, we construct $DC_r(X, Y)$, displayed in Fig. 5, for $r = 0.000109$ and map the trajectories to edge-paths as before. Classifying the trajectories using our approach, we obtain 24 trajectory classes as color-coded in Fig. 6. Some of these classes are separately displayed in the bottom part of that figure. As is visible in the figure, we are able to cleanly separate interesting driver behaviors with our approach - note for example the clover leaf trajectory class, where a driver crossed several bridges before continuing along the same driving direction.

### B. Incorporation of probability densities

In this experiment, we illustrate the use of a probabilistic model in conjunction with our approach. Consider the indoor scene in Fig. 7. We would like to model the behavior of pedestrians in this space using a dataset of 197 trajectories with 26029 datapoints $X \subset \mathbb{R}^2$ from [2]. We extract 45

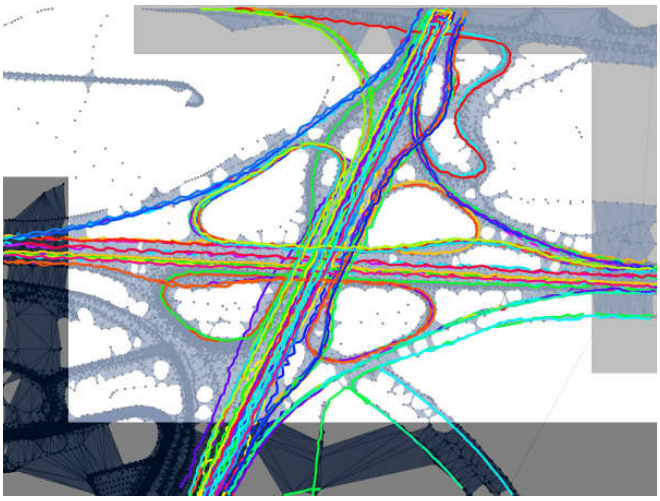

Fig. 5. OpenStreetMap [25] dataset of GPS traces intersecting a bounding box around a highway crossing. We display the discretization of those trajectories traveling between the indicated gray regions as well as the complex $DC_r(X, Y)$, $r = 0.000109$. Note that some of the scattered points in the complex are due to GPS datapoints of other trajectories in the bounding box which did not travel between the shaded regions.

trajectories traversing between the shaded regions, displayed in black on the left. Besides these trajectories, we consider a Gaussian Mixture Model $M$, modeling the positions of potential collision threats in this space (shown in color). We construct a simplicial complex filtration $P(X, Y)$ starting with the sub-complex $A(Y) \subset D(X)$ at filtration zero, as before, but inserting simplices $\sigma = [v_0, \ldots, v_p]$ of $D(X)$ in order of $max_{i=0}^{p} p(v_i)$, where $p : \mathbb{R}^2 \to \mathbb{R}$ denotes the probability density of the mixture model. The first persistence diagram for this filtration has three large persistence intervals: one corresponding to each mode of $p$, with the middle Gaussian resulting in the global maximum of $p$. All displayed trajectories have vertices lying in the complex $P_r(X, Y)$, for $r = 0.5$. At $r = 0.59$, we obtain 4 trajectory classes as indicated in the left figure, but as $r$ increases these classes start to merge, starting with the red and black classes which are separated by a rather small mode of $p$. Our classification scheme in this case hence allows us to understand the movement of pedestrians *relative* to the modes of a mixture model.

## V. Conclusion

We have proposed a versatile topological trajectory clustering approach for trajectories with initial and terminal points in a specified region $R$. While our approach is applicable in arbitrary dimensions, we have focused on experiments in 2D, showing that our approach can extract interesting motion classes from large real-world trajectory datasets. In future work, we would like to focus on investigating our classification procedure for higher dimensional trajectory data, arising in computer animation and robotics. We believe that the approach presented here provides a classification approach which is complementary to classical probabilistic

Preprint submitted to 2016 IEEE International Conference on Robotics and Automation. Received September 15, 2015.

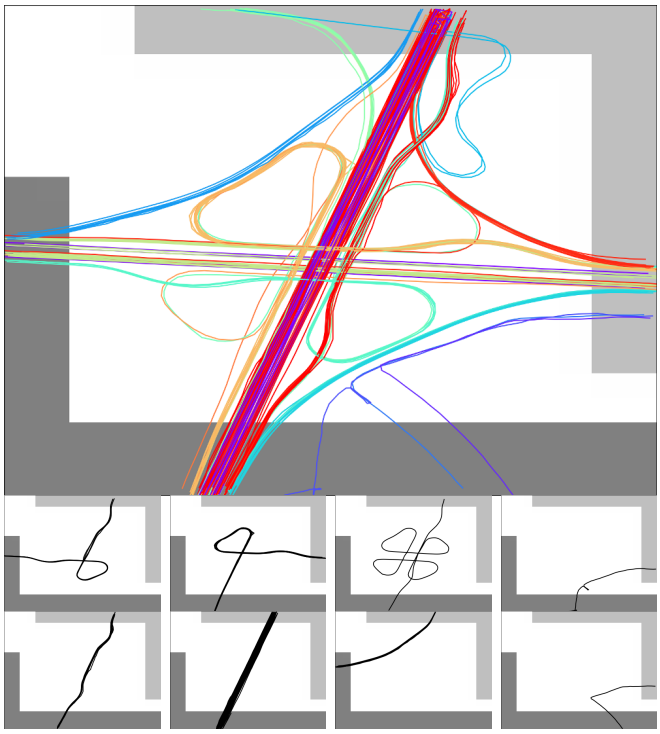

Fig. 6. We display a dataset of GPS car traces near a highway crossing for trajectories which travel between the indicated gray regions. The 24 found topological classes using $DC_r(X,Y)$ at $r = 0.000109$ are indicated by color and selected classes are separately displayed in the small figures below.

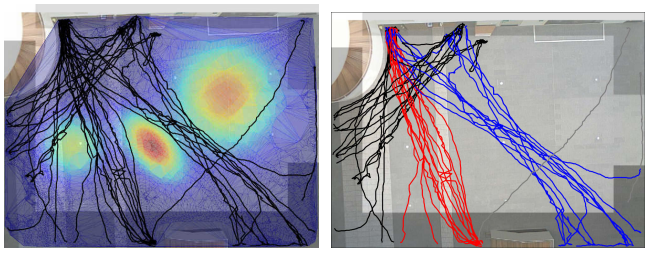

Fig. 7. Subset of 45 traces of pedestrian data from [2]. On the right, we display the resulting classification based on the probabilistic collision model shown on the left (in color) and at filtration value of $r = 0.59$.

or distance based clustering techniques, due to its scalability to very large datasets and since the methods presented here can extract global topological information about data that main-stream machine learning techniques have so far not been able to benefit from.

## REFERENCES


[1] B. D. Argall, S. Chernova, M. Veloso, and B. Browning, "A survey of robot learning from demonstration," *Robotics and autonomous systems*, vol. 57, no. 5, pp. 469–483, 2009.

[2] B. Fisher, "Edinburgh informatics forum pedestrian database," http://homepages.inf.ed.ac.uk/rbf/FORUMTRACKING/, 2015.

[3] U. Bauer and H. Edelsbrunner, "The morse theory of Čech and Delaunay filtrations," in *Proc. of the Thirtieth Annual Symp. on Comp. Geometry*, ser. SOCG'14. New York, NY, USA: ACM, 2014, pp. 484:484–484:490.

[4] U. Bauer, M. Kerber, and J. Reininghaus, "PHAT (Persistent Homology Algorithm Toolbox)," http://code.google.com/p/phat/.

[5] A. Billard, S. Calinon, R. Dillmann, and S. Schaal, "Robot programming by demonstration," in *Springer handbook of robotics*. Springer, 2008, pp. 1371–1394.

[6] L. Brun, A. Saggese, and M. Vento, "A clustering algorithm of trajectories for behaviour understanding based on string kernels," in *Signal Image Technology and Internet Based Systems (SITIS), 2012 Eighth International Conference on*. IEEE, 2012, pp. 267–274.

[7] K. Buchin, M. Buchin, J. Gudmundsson, M. Löffler, and J. Luo, "Detecting commuting patterns by clustering subtrajectories," *International Journal of Computational Geometry & Applications*, vol. 21, no. 03, pp. 253–282, 2011.

[8] G. Carlsson, "Topology and data," *Bull. Amer. Math. Soc. (N.S.)*, vol. 46, no. 2, pp. 255–308, 2009.

[9] CMU Graphics Lab, "CMU graphics lab motion capture database," http://mocap.cs.cmu.edu/, 2015.

[10] V. De Silva, D. Morozov, and M. Vejdemo-Johansson, "Dualities in persistent (co) homology," *Inverse Problems*, vol. 27, no. 12, p. 124003, 2011.

[11] H. Edelsbrunner and J. Harer, "Persistent homology-a survey," *Contemporary mathematics*, vol. 453, pp. 257–282, 2008.

[12] H. Edelsbrunner and J. L. Harer, *Computational topology: an introduction*. AMS Bookstore, 2010.

[13] A. Fod, M. J. Matarić, and O. C. Jenkins, "Automated derivation of primitives for movement classification," *Autonomous robots*, vol. 12, no. 1, pp. 39–54, 2002.

[14] S. Gaffney and P. Smyth, "Trajectory clustering with mixtures of regression models," in *Proceedings of the fifth ACM SIGKDD int. conf. on Knowledge discovery and data mining*. ACM, 1999, pp. 63–72.

[15] K. Goldberg and B. Kehoe, "Cloud robotics and automation: A survey of related work," *EECS Department, University of California, Berkeley, Tech. Rep. UCB/EECS-2013-5*, 2013.

[16] A. Hatcher, *Algebraic topology*. Cambridge: Cambridge University Press, 2002.

[17] ICML, "Workshop: Topological methods for machine learning," 2014.

[18] O. C. Jenkins and M. Mataric, "Automated derivation of behavior vocabularies for autonomous humanoid motion," in *Proceedings of the second international joint conference on Autonomous agents and multiagent systems*. ACM, 2003, pp. 225–232.

[19] K. K., D. L., and I. E., "Gaussian process regression flow for analysis of motion trajectories," in *IEEE ICCV*. IEEE Computer Society, November 2011.

[20] R. A. Knepper, S. S. Srinivasa, and M. T. Mason, "Toward a deeper understanding of motion alternatives via an equivalence relation on local paths," *Int. Journal of Robotics Research*, vol. 31, no. 2, pp. 167–186, 2012.

[21] J.-G. Lee, J. Han, and K.-Y. Whang, "Trajectory clustering: a partition-and-group framework," in *Proceedings of the 2007 ACM SIGMOD int. conf. on Management of data*. ACM, 2007, pp. 593–604.

[22] MarineTraffic, "Marinetraffic database," http://marinetraffic.com/, 2014.

[23] N. Nguyen, D. Phung, S. Venkatesh, and H. Bui, "Learning and detecting activities from movement trajectories using the hierarchical hidden markov model," in *CVPR 2005*, vol. 2. IEEE, 2005, pp. 955–960.

[24] NIPS, "Workshop: Algebraic topology and machine learning," 2012.

[25] OpenStreetMap, "Open Street Map," http://www.openstreetmap.org, 2015.

[26] P. Pastor, H. Hoffmann, T. Asfour, and S. Schaal, "Learning and generalization of motor skills by learning from demonstration," in *Robotics and Automation, 2009. ICRA'09. IEEE International Conference on*. IEEE, 2009, pp. 763–768.

[27] C. Piciarelli, C. Micheloni, and G. L. Foresti, "Trajectory-based anomalous event detection," *Circuits and Systems for Video Technology, IEEE Trans.*, vol. 18, no. 11, pp. 1544–1554, 2008.

[28] F. T. Pokorny, M. Hawasly, and S. Ramamoorthy, "Multiscale topological trajectory classification with persistent homology," in *Proceedings of Robotics: Science and Systems*, Berkeley, USA, July 2014.

[29] ——, "Topological trajectory classification with filtrations of simplicial complexes and persistent homology," *IJRR*, 2015.

[30] C. Saunders, D. R. Hardoon, J. Shawe-Taylor, and G. Widmer, "Using string kernels to identify famous performers from their playing style," in *Machine Learning: ECML 2004*. Springer, 2004, pp. 384–395.

[31] H. Wang, A. Klaser, C. Schmid, and C.-L. Liu, "Action recognition by dense trajectories," in *CVPR*. IEEE, 2011, pp. 3169–3176.

[32] Y. Zheng and X. Zhou, *Computing with spatial trajectories*. Springer, 2011.