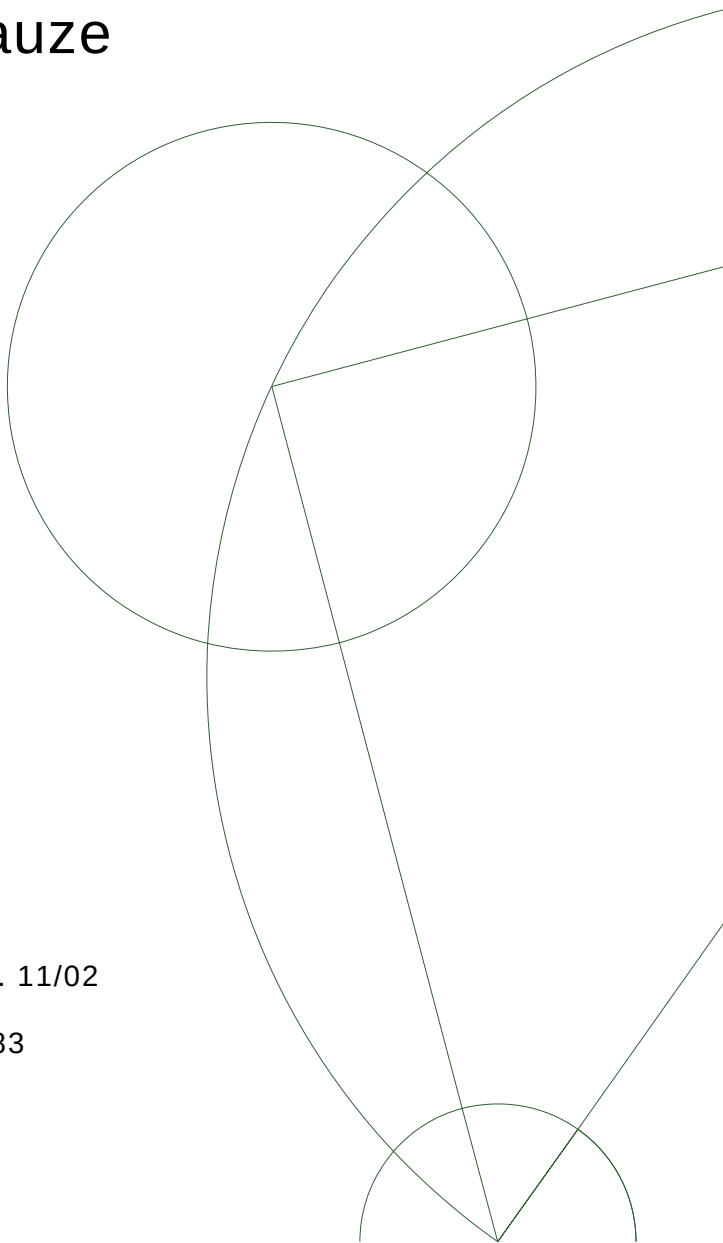
DEPARTMENT OF COMPUTER SCIENCE (DIKU)
UNIVERSITY OF COPENHAGEN

A geometric framework for statistics on trees

Aasa Feragen, Mads Nielsen, Søren Hauberg, Pechin Lo, Marleen de Bruijne and François Lauze



April 2011

Technical Report no. 11/02

ISSN: 0107-8283

Contents

Abstract

This is a technical report summarizing recent results on a statistical shape analysis framework for tree-like shapes. A theoretical approach to the space of tree-like shapes and how its geometry can be utilized for tree-shape statistics is found in Chapter 1. In Chapter 2 we discuss several approaches to computing average tree-shapes.

Chapter 1

Geometry and statistics in spaces of tree-like shapes

In order to develop statistical methods for shapes with a tree-structure, we construct a shape space framework for tree-like shapes and study metrics on the shape space. The shape space has singularities, which correspond to topological transitions in the represented trees. We study two closely related metrics, TED and QED. The QED is a quotient euclidean distance arising from the new shape space formulation, while TED is the classical tree edit distance. Using Gromov's metric geometry we gain new insight into the geometries defined by TED and QED. In particular, we show that the new metric QED has nice geometric properties which facilitate statistical analysis, such as existence and local uniqueness of geodesics and averages. TED, on the other hand, has algorithmic advantages, while it does not share the geometric strongpoints of QED. We provide a theoretical framework as well as computational results such as matching of airway trees from pulmonary CT scans and geodesics between synthetic data trees illustrating the dynamic and geometric properties of the QED metric.

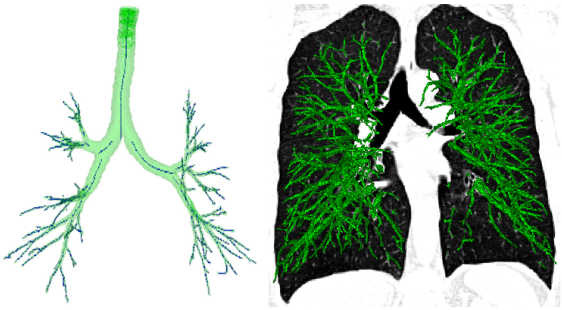


Figure 1.1: Examples of anatomical tree-like structures are airways and blood vessels in lungs [PAD+09].

1.1 Introduction

Trees are fundamental structures in nature, where they appear as delivery systems for air and fluids, as skeletal structures, or describing hierarchies. Examples encountered in image analysis and computational biology are airway trees [TMP+05, KKNN06, MKS+09, BLWH06, vGBvR08], vascular systems [CAM+05, CFM01, MBK+09, SBK+10], shock graphs [SKK04, BWL+09], scale space hierarchies [KF05, DPS+09] and phylogenetic trees [BHV01, OP11].

In the past 20 years, extensive work has been done related to the comparison of such structures in terms of matching [SBK+10, CAM+05, KSDD03, MKS+09, TMP+05], object recognition [DPS+09, KSK01] and machine learning [RB09b, FVS+10, JO09, JO10] based on inter-tree distances. However, the existing tree-distance frameworks are algorithmic rather than geometric. As we shall see, this yields problems even in the basic problem of finding the average of two tree-shapes. There exists no principled approach to studying the space of tree-structured data, and as a consequence, attempts to find average trees or perform a modes of variation analysis are not well-defined. Statistical methods for tree-structured data would have endless applications. For instance, in medical image analysis one could better study the documented changes in airway-shape geometry and structure in COPD patients [WDE+09, SLD+11] in order to improve tools for computed-aided diagnosis and prognosis.

In this paper we study shapes with a tree-like structure, represented as sets of embedded curves connected by a hierarchical rooted tree. In analogy with the well-known statistical shape analysis for curves and surfaces, we define a coherent space of tree-like shapes, and give it a metric, called Quotient Euclidean Distance (QED), where distances are realized as lengths of geodesic deformations.

One of our goals is to be able to compute means of sets of tree-like shapes. In as simple settings as two-point datasets, means are closely connected to geodesics, in the sense that the midpoint of a geodesic from a to b will be a mean for the dataset $\{a, b\}$. If there is more than one geodesic connecting a to b, with different midpoints, then there will also be more than one mean describing the dataset. Because of this, (local) uniqueness of geodesics is crucial to having well-defined means. Using methods from Gromov's metric geometry, which are novel in computer vision and pattern recognition, we show that locally, geodesics are unique, paving the road for further statistical analysis. We show that in the local situation, we can uniquely define several concepts of average tree.

Moreover, we see how a natural variation of the QED metric on the defined shape space gives rise to the classical Tree Edit Distance (TED) metric, and through an analysis of the tree-space geometry we see how TED does not give the same natural statistical properties as QED. This explains why using TED for computing mean trees must always be accompanied by a carefully engineered choice of edit paths in order to give well-defined results; choices which can yield average trees which are substantially different from the trees in the dataset [TK10].

In short, our paper provides a principled approach to tree-shape analysis, based on a shape space construction for tree-like shapes. Within this shape space, we are able to naturally define and compare the two natural tree-shape metrics TED and QED, and prove that QED has some very desirable properties

for statistical analysis, that TED does not share. As a result, we are able to investigate statistical methods for tree-like structures which have previously not been possible, such as different concepts of average tree.

1.1.1 A space of tree-like shapes

In a continuous space of tree-like shapes we should be able to continuously deform any given tree-shape into any other by traveling along a path in tree-shape space that connects the two shapes. Holding on to this idea, we can get an intuitive idea of how the tree-space should be connected.

First consider the space of all rooted tree-shapes with a given tree-topological structure $\mathscr{T} = (V, E, r)$. Assume that the edges E in the tree have a fixed numbering, and that the edge shapes are described by a vector in $\mathbb{R}^N$, for instance consisting of a number of landmark points. Now, any such tree can be described by a long vector in $\prod_{e\in E} \mathbb{R}^N$.

Next, consider the space of all tree-shapes in the whole entire world. This set can be partitioned into an infinite set $\mathcal{C}$ of classes C, where each class contains all tree-shapes with a given tree-topological structure $\mathscr{T}_C = (V_C, E_C, r_C)$. Now the tree-space is a disjoint union $X = \bigsqcup_{C\in\mathcal{C}} \prod_{e\in E_C} \mathbb{R}^N$, and every tree-shape you can come up with is represented by a point in X.

However, there is a problem with this space X. Consider a class C whose tree-shapes all have the same tree topology $\mathscr{T}$, and consider a sequence of trees in C where a certain edge is becoming smaller and smaller, converging towards the edge described by a zero vector. Geometrically, this means that the edge is disappearing, and we are approaching a tree-shape with a different tree topology. We can similarly approach the same tree-shape through sequences of trees with a third tree topology, and so on. This thought experiment intuitively illustrates how the different components of X should be glued together in order to form a connected tree-shape space.

Note, in particular, that the components are glued together along a subspace in which a new tree topology is found, which is different from the topology of the trees in the different sequences. In other words, the tree-space is partitioned into components of different dimension, in which trees with different topological structures are found, which intersect each other in yet new components – see fig. 1.2a for an illustration. As seen in the figure, a path in tree-space corresponds to a tree-shape deformation, and passing through a lower-dimensional component corresponds to an internal structural transition in the deforming tree. These structural transitions are key to defining geodesics between the tree-shapes, as we shall do later on in the paper.

The main theoretical contributions of the paper are found in sections 1.3 and 1.4. The formal construction of the tree-shape space is the topic of section 1.3, whose main result is theorems 1.4 and 1.6. These theorems state that we can, indeed, form a geometric tree-shape space and endow it with two natural metrics which can both be realized as lengths of geodesic tree-shape deformations. Moreover, we see in section 1.3.5 that one of this metrics is the classical TED. This illustrates, in particular, that the relatively complicated construction of the tree-shape space is not a hindrance for finding intuitive algorithms for tree-distance computations. In theorems 1.7 and 1.16 we show that the second metric, QED, has geometric properties that make it particularly well suited for statistical analysis. TED does not share these properties.

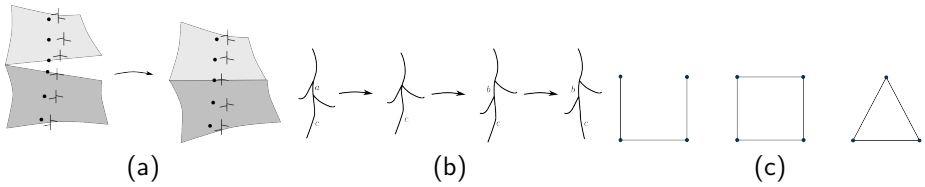


Figure 1.2: (a) Tree-shapes with different topology live in different components of tree-space. These are glued together along subsets with collapsed versions of the tree-shapes. (b) The TED moves: Remove an edge a, add an edge b, deform an edge c. (c) In graph representations [JO09], zero branch descriptors denote "missing" branches (left), whereas in our tree representation they describe collapsed branches, (right). This induces very different spaces of graphs/trees.

The paper is organized as follows: In section 1.2 we give an overview of related work. The tree-space is defined in section 1.3, and its statististical properties are analyzed in section 1.4. Using Gromov's approach to metric geometry [Gro87] we gain insight into the geometric properties of two different metrics; one which is essentially Tree Edit Distance (TED) and one which is a quotient euclidean distance (QED). We pay particular respect to the properties of geodesics and averages, which are essential for statistical shape analysis. In section 1.5 we discuss how to overcome the computational complexity of both metrics. A simple QED approximtion is discussed in Section 1.6, and in Section 1.7 we illustrate the properties of QED both by computing geodesics and means for synthetic planar data trees as well as by computing QED distances for sets of $3D$ pulmonary airway trees which show promising tendencies for classification based on the QED metric.

1.2 Related work

Metrics on sets of tree-structured data have been studied by different research communities for the past 20 years. The best-known approach in the computer vision community is perhaps Tree Edit Distance (TED), which has been used extensively for shape matching and recognition based on medial axes and shock graphs [SKK04, KTSK00, SKK02, KSK01, TK10, TH03]. TED and more generally graph edit distance have also been popular in the pattern recognition community, and are still used for most modern distance-based pattern recognition approaches to trees and graphs [FVS$^+$10, RB09b, RB09a]. One of our main goals is, however, to develop a framework analogous to the classical shape statistics [Ken84, HHM10] to tree-like shapes; particularly in order to define and compute average tree-shapes and analyze modes of variations in a style similar to PCA.

In TED, the distance between two trees is the minimal total cost of deforming the first tree into the second, using tree basic operations: remove an edge, add an edge, and deform an edge, see fig. 1.2b. The cost of changing an edge into another is, in this paper, the Euclidean distance between their landmark point vectors, both aligned so that the first landmark point is at the origin. Removing an edge is equivalent to deforming it to a collapsed edge, and adding an edge is the inverse of removal. It is easy to see that the TED metric will nearly always

have infinitely many edit paths, or geodesics, between two given trees.

As a result, even the problem of finding the average of two trees is no longer well posed. Without local uniqueness of geodesics it becomes hard to meaningfully define and compute average shapes or modes of variation. This problem can be solved to some extent by choosing a preferred edit path [FVS$^+$10, RB09b], but there will always be a risk of having chosen the "wrong" edit path. Trinh and Kimia [TK10] have recently used TED for computing average medial axes using the simplest possible edit paths, leading to average shapes which are substantially different from most of the dataset shapes.

Statistics on tree-structured objects have recently sparked a growing interest in the statistical community. Wang and Marron [WM07] have studied metric spaces of trees and defined a notion of average tree called the median-mean as well as a version of PCA, which finds modes of variation in terms of so called *tree-lines*, encoding the maximum amount of structural and attributal variation. Aydin et al. [APW$^+$09] extend this work by finding efficient algorithms for computing the PCA. However, the metric defined by Wang and Marron does not define a natural geodesic structure on the space of trees with varying topological structure, as it places a very large emphasis on the tree-topological structure of the trees. More precisely, the metric has discontinuities in the sense that a sequence of trees with a shrinking branch will not converge to a tree that does not have that branch. Wang and Marron analyze datasets consisting of brain blood vessels, which are trees with few, long branches, and for such data the strong emphasis on tree topology might work well. However, this metric is not suitable for studying large trees with topological variations and noise, such as airways or most other tree-structures appearing in medical images, as the metric punishes structural changes much harder than shape variation with constant tree-topological structure. Likewise, their PCA mainly encodes combinatorics, whereas in a continuous space of attributed trees we would like the interplay between tree topology and attributes (in our case, branch geometry) to be stronger.

A rather different approach is that of Jain and Obermayer [JO09, JO10], who define metrics on attributed graphs. The graphs are represented using incidence matrices, and the space of graphs is defined as the quotient of the Euclidean space of incidence matrices by the group of vertex relabelings. The graph-space inherits a metric from Euclidean space, which gives it the structure of an orbifold. Means are computed using Lipschitz analysis, giving fast computations. The graph-space construction by Jain and Obermayer is similar to the tree-space presented in this paper in the sense that both spaces are constructed as quotients of a Euclidean space; however, there are some modeling differences that result in significant differences in the graph- and tree-space geometries. In an incidence matrix M representing a graph the entry M_{ij} is zero if there is no edge connecting the ith and jth vertices. As a consequence, such a model cannot describe what happens when an edge is shrinking to finally disappear, and as discussed above, such a model is not suitable for most medical image analysis applications. In the tree-space presented here, we use zero edge attributes to represent collapsed edges. As a result, the tree-space quotient will make identifications that are not represented by a group acting on vertices or edges, and the tree-space presented in this article will not be an orbifold.

Tree-like structures also appear in other settings, for instance as phylogenetic trees describing genetics. Billera et al. [BHV01] have created a mathe-

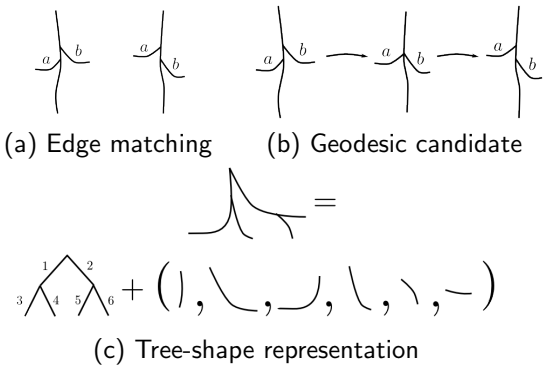


Figure 1.3: (a-b) A good metric must handle edge matchings which are inconsistent with tree topology. (c) Tree-like shapes are encoded by an ordered binary tree and a set of attributes describing edge shape.

matical framework for geodesic analysis of phylogenetic trees, and Owen and Provan [OP11] have developed efficient algorithms for computing geodesics in the space of phylogenetic trees.

We have previously [FLN10, FLL$^+$10] studied geodesics between small tree-shapes in the same type of singular shape space as studied here. In this paper we give in-depth explanations and more extensive proof-of-concept examples illustrating the possibilities in the QED metric.

1.3 The space of tree-like shapes

Before defining a tree-space and giving it a geometric structure, let us discuss which properties are desirable for a tree-shape model. In order to compute average trees and analyze variation in datasets, we require, at the very least, local existence and uniqueness properties for geodesics. As discussed in the previous section, if we do not have (locally) unique geodesics, then we do not even have well-defined means for datasets with two trees. When geodesics exist, we want the topological structure of the intermediate trees to reflect the resemblance in structure of the trees being compared – in particular, *a geodesic passing through the trivial one-vertex tree should indicate that the trees being compared are maximally different.* Perhaps more important, we would like to compare trees where edge matching is inconsistent with tree topology, as in fig. 1.3a; specifically, we would like to find geodesic deformations in which the tree topology changes when we have such edge matchings, for instance as in fig. 1.3b.

1.3.1 Representation of trees

In this paper, a "tree-shape" is an embedded tree in $\mathbb{R}^2$ or $\mathbb{R}^3$. More precisely, a tree-shape will consist of a series of edge embeddings, glued together as defined by a rooted combinatorial tree. Tree-shapes are invariant to translation, but we do not remove scale and rotation from the definition of a tree-shape.

Any tree-like (pre-)shape is represented as a pair $(\mathscr{T}, x)$ consisting of a rooted, planar, binary tree $\mathscr{T} = (V, E, r)$ with edge attributes. Here, $\mathscr{T}$ de-

scribes the tree topology, and the attributes describe edge geometry, as illustrated in fig. 1.3c. The attributes are represented by a point $x \in \prod_{e \in E} A$, where A is the attribute space. Tree-shapes which are *not* binary are represented by the binary tree $\mathscr{T}$ in a very natural way by allowing constant, or collapsed, edges, represented by the zero scalar or vector attribute. Endowing an internal edge with a zero attribute corresponds to collapsing that edge, and in this way *an arbitrary attributed tree can be represented as an attributed* **binary** *tree*, see fig. 1.4b.

Representing tree-shapes by binary trees is natural for several reasons. Binary trees are *stable* in the sense that whenever a perturbation of a binary tree-shape is sufficiently small, the topological structure of the tree-shape remains unchanged. Conversely, a trifurcation or higher-order vertex can always be turned into a series of bifurcations sitting close together by an arbitrarily small perturbation. In our representation, thus, trifurcations are represented as two bifurcations sitting infinitely close together, etc. In the case of airway trees, which are of particular interest for us, binary tree-shape representations are in agreement with Weibel's dichotomous airway tree model [Wei09].

The edge attributes in A describe edge shape, and could, e.g., be edge length, landmark points or edge parametrizations. In this work, we use open curves translated to start at the origin, described by a fixed number of landmark points. Thus, throughout the paper, the attribute space A is $(\mathbb{R}^d)^n$ where $d = 2$ or 3 and n is the number of landmark points per edge. Collapsed edges are represented by a sequence of origin points.

In order to compare trees of different sizes and structures, we need to represent them in a unified way. We describe all shapes using the *same* binary tree $\mathscr{T}$ to encode tree topology. By choosing a sufficiently large binary tree we can represent all the trees in our dataset by filling out with collapsed edges. We call $\mathscr{T}$ *the maximal binary tree.*

Trees embedded in the plane are given a natural edge order induced by the left-right order on the children of any edge. Similarly, a combinatorial tree whose edges are ordered will always have a unique, implicitly defined embedding in the plane such that the order of siblings ascends from left to right. For this reason we use the terms "planar tree" and "ordered tree" interchangingly. We initially study metrics on the set of ordered binary trees; later we use them to compute distances between unordered trees by considering all possible orders. This leads to potential computational challenges, which are discussed in Section 1.5.

Fix an ordered maximal binary tree $\mathscr{T}$ with edges E, which encodes the connectivity of all our trees. Any attributed tree T is now represented by a point $x = (x_e)_{e \in E}$ in $X = \prod_{e \in E} (\mathbb{R}^d)^n$, where the coordinate x_e describes the shape of the edge x. We call X the tree pre-shape space, since some tree-shapes are represented by several points in X. This causes problems as some natural tree-deformations are not represented as continuous paths in X, see fig. 1.4a. We solve this problem by inducing a refined space, which will become the space of tree-like shapes.

1.3.2 The singular space of ordered tree-shapes

We go from pre-shapes to shapes by identifying those pre-shapes which define the same shape.

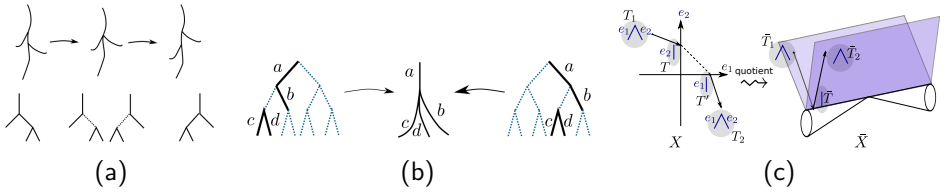


Figure 1.4: (a) The tree deformation shown in the top row does not correspond to a path in X, as the two representations of the intermediate tree are found at distinct points in X. (b) Higher-degree vertices are represented by collapsing internal edges (dotted line = zero attribute = collapsed edge). We identify those tree pre-shapes whose collapsed structures are identical: here x_1 and x_2 represent the same tree T. (c) The simplest non-trivial tree-shape space, consisting of trees with two branches endowed with scalar attributes. Along the x- and y-axes we find trees with a single branch; hence, for each real number a the tree-shape found at $T' = (a, 0)$ is also represented at $T = (0, a)$. We build the tree-shape space by gluing the different representations of the same tree-shapes, such as the representations T and T', together, obtaining the shape space shown on the right. Note also the path from $\bar{T}_1$ to $\bar{T}_2$ through $\bar{T}$ in $\bar{X}$ on the right. The corresponding path in the pre-shape space X involves a "teleportation" from the representation T to the representation T'.

Consider two ordered tree-shapes where internal edges are collapsed, and replace their binary representations by collapsed representations where the zero attributed edges have been removed. The depth-first orders of the original trees induce well-defined depth-first orders on the collapsed trees. We say that two ordered tree-shapes are *the same* when their collapsed ordered topological structures are identical, and the edge attributes on corresponding non-collapsed edges are identical as well, as in fig. 1.4b. Thus, tree identifications come with an inherent bijection of subsets of E: If we identify $x, y \in X = (\mathbb{R}^N)^E$, denote $E_1 = \{e \in E | x_e \neq 0\}, E_2 = \{e \in E | y_e \neq 0\}$; the identification comes with an order preserving bijection $\varphi \colon E_1 \to E_2$ identifying those edges that correspond to the same edge in the collapsed tree-shape. Note that φ will also correspond to similar equivalences of pairs of trees with the same topology, but other attributes.

The bijection φ next induces a bijection $\Phi \colon V_1 \to V_2$ given by $\Phi \colon (x_e) \mapsto (x_{\varphi(e)})$. Here, $V_1 = \{x \in X | x_e = 0 \text{ if } e \notin E_1\}$ and $V_2 = \{x \in X | x_e = 0 \text{ if } e \notin E_2\}$ are subspaces of X where, except for at the axes, the topological tree structure is constant, and for $x \in V_1$, $\Phi(x) \in V_2$ describes the same shape as x.

We define a map Φ for each pair of identified tree-structures, and form an equivalence on X by setting $x \sim \Phi(x)$ for all x and Φ. For each $x \in X$ we denote by $\bar{x}$ the equivalence class $\{x' \in X | x' \sim x\}$. The quotient space $\bar{X} = (X/\sim) = \{\bar{x} | x \in X\}$ of equivalence classes $\bar{x}$ is the space of tree-like shapes.

Quotient spaces are standard constructions from topology and geometry, where they are used to glue spaces together [BH99, chapter 1.5]. The geometric interpretation of the identification in the tree-space quotient is that we are folding and gluing the pre-shape space space along the identified subspaces; i.e. when $x_1 \sim x_2$ we glue the two points x_1 and x_2 together. See the simple tree-shape space in fig. 1.4c for an intuitive illustration.

1.3.3 Metrics on the space of ordered trees

Given a metric d on the Euclidean pre-shape-space $X = \prod_{e \in E} \mathbb{R}^N$ we define the standard quotient pseudometric [BH99] $\bar{d}$ on the quotient space $\bar{X} = X/\sim$ by setting

$$\bar{d}(\bar{x}, \bar{y}) = \inf \left\{ \sum_{i=1}^{k} d(x_i, y_i) | x_1 \in \bar{x}, y_i \sim x_{i+1}, y_k \in \bar{y} \right\}. \tag{1.1}$$

This amounts to finding the optimal path from $\bar{x}$ to $\bar{y}$, consisting of any number k of concatenated Euclidean lines, passing through $k-1$ identified subspaces, as shown in fig. 1.4c. It is clear from the definition that the distance function $\bar{d}$ is symmetric and transitive. It is, however, defined as an infimum of a set of real numbers, giving a risk that the distance between two distinct tree-shapes is zero, as occurs with some intuitive shape distance functions [MM04]. This is why d is called a *pseudo*metric. We need to prove that it actually is a metric; i.e., that $\bar{d}(\bar{x}, \bar{y}) = 0$ implies $\bar{x} = \bar{y}$.

We define two metrics on X, which come from two different ways of combining the individual edge distances: The metrics d_1 and d_2 on $X = \prod_{e \in E} \mathbb{R}^N$ are induced by the norms

$$\|x - y\|_1 = \sum_{e \in E} \|x_e - y_e\|, \tag{1.2}$$

$$\|x - y\|_2 = \sqrt{\sum_{e \in E} \|x_e - y_e\|^2}. \tag{1.3}$$

From now on, d and $\bar{d}$ will denote either the distance functions d_1 and $\bar{d}_1$, or d_2 and $\bar{d}_2$. Recall that it is not yet clear that the $\bar{d}_i$ are going to define metrics on $\bar{X}$.

We shall prove the following:

Theorem 1.4 *The distance function $\bar{d}$ is a metric on $\bar{X}$, which is a contractible, complete, proper geodesic space.*

Proof. See Appendix 1.9. □

This means, in particular, that given any two trees, we can always find a geodesic between them in both metrics $\bar{d}_1$ and $\bar{d}_2$.

Remark 1.5 It can be shown that for *any* metric d on X, the induced pseudometric $\bar{d}$ on $\bar{X}$ is a metric.

1.3.4 From planar trees to spatial trees

The world is not two-dimensional, and for most applications it is necessary to study embedded trees in $\mathbb{R}^3$. As far as attributes in terms of edge embeddings and landmarks are concerned, this is, however, not very different from the planar embedded trees – the main difference from the $\mathbb{R}^2$ case comes from the fact that trees in $\mathbb{R}^3$ have no canonical edge order. The left-right order on children of planar trees gives an implicit preference for edge matchings, and hence reduces the number of possible matches. When we no longer have this preference, we

generally need to consider all orderings of the same tree and choose the one which minimizes the inter-tree distance to a certain fixed, ordered tree.

We define the space of spatial tree-like shapes as the quotient $\bar{\bar{X}} = \bar{X}/G$, where G is the group of reorderings of the standard tree. The metric $\bar{d}$ on $\bar{X}$ induces a quotient pseudometric $\bar{\bar{d}}$ on $\bar{\bar{X}}$.

Note that G is a finite group, which means that $\bar{\bar{X}}$ is locally well-behaved almost everywhere. In particular, off fixed-points for the action of G on $\bar{X}$, the projection $\bar{p}\colon \bar{X} \to \bar{\bar{X}}$ is a local isometry; hence the geometry from $\bar{X}$ is preserved off the fixed points. Geometrically, a fixed point is a tree-shape where a reordering of certain branches does not change the tree-shape; that is, some pair of sibling edges must have the same shape attributes. In particular, the fixed points are non-generic because they belong to the lower-dimensional subset of $\bar{X}$ where the two (sibling) edges have identical attributes. Using standard results on compact transformation groups along with similar techniques as for theorem 1.4, we can prove:

Theorem 1.6 *For $\bar{\bar{d}}$ induced by either $\bar{d}_1$ or $\bar{d}_2$, the function $\bar{\bar{d}}$ is a metric and the space $(\bar{\bar{X}}, \bar{\bar{d}})$ is a contractible, complete, proper geodesic space.*

While considering all different possible orderings of the tree makes perfect sense from the geometric point of view, in reality this becomes an impossible task as the size of the trees grow beyond a few generations. In real applications we can, however, efficiently reduce complexity using a few tricks. These tricks include taking tree- and tree-space geometry into account as discussed in Section 1.5, but also inducing an order on unordered trees by aligning them with one fixed tree. This corresponds to using the local isometry property of the projection $\bar{p}\colon \bar{X} \to \bar{\bar{X}}$, which is similar to the use of horizontal geodesics e.g. used by Huckemann et al. [HHM10].

1.3.5 Geometric interpretation of the metrics

It is easy to see from the definition that the metrics $\bar{d}_1$ and $\bar{\bar{d}}_1$ coincide with the classical tree edit distance (TED) metric for ordered and unordered trees, respectively. This shows that the mathematical construction of tree-space is not an obstruction to finding intuitive algorithms for inter-tree distances.

The metrics $\bar{d}_2$ and $\bar{\bar{d}}_2$ are descents of the Euclidean metric on $\bar{X}$, and geodesics in this metric are concatenations of straight lines in flat regions. We call them the QED metric on ordered and unordered trees, respectively, for quotient euclidean distance.

In Section 1.4.5 we compare the two metrics using example geodesic deformations.

1.4 Curvature in the space of tree-like shapes

Uniqueness of geodesics and means is closely connected to the curvature of the dataspace, which is the topic of the current section. Using methods from metric geometry [Gro87] we shall investigate the curvature of the tree-shape space with the QED and TED metrics.

The next theorem states that in the tree-shape space endowed with the QED metric, any randomly selected point has a corresponding radius within which

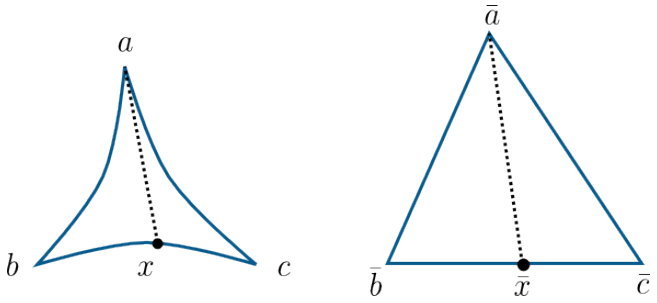


Figure 1.5: A metric space is a $CAT(0)$ space if, for a geodesic triangle abc and for any point x on the triangle, the distance from x to the opposite vertex is at most as long as the corresponding distance in the planar comparison triangle $\bar{a}\bar{b}\bar{c}$.

the tree-space has non-positive curvature. Later in this section we shall use this fact to show that datasets within that same radius have unique averages.

Theorem 1.7 *i) Endow $\bar{X}$ with the QED metric $\bar{d}_2$. A generic point $\bar{x} \in \bar{X}$ has a neighborhood $U \subset \bar{X}$ in which the curvature is non-positive. At non-generic points, the curvature of $(\bar{X}, \bar{d}_2)$ is unbounded.*

ii) Endow $\bar{X}$ with the TED metric $\bar{d}_1$. The metric space $(\bar{X}, \bar{d}_1)$ does not have locally unique geodesics anywhere, and the curvature of $(\bar{X}, \bar{d}_1)$ is everywhere unbounded.

Formally, a generic property is a property which holds on an open, dense subset. One interpretation of this, given a probability measure compatible with the topology, is that generic properties hold with probability one. Thus, for a random tree-shape, we can safely assume that tree-space is locally non-positively curved at that shape.

Similarly, by a non-generic property we mean a property whose "not happening" is generic – i.e. a property that we need not worry about for a random tree-shape.

In order to understand and prove theorem 1.7, we need the concept of curvature in metric spaces. In spite of its simplicity and elegance, this concept from metric geometry is novel in computer vision, and we shall spend a little time introducing it.

1.4.1 Curvature in metric spaces

Since general metric spaces can have all kinds of anomalies, the concept of curvature in such spaces is usually defined though a comparison with spaces that we understand well. More precisely, the metric spaces are studied using *geodesic triangles*, which are compared with corresponding *comparison triangles* in model spaces with a fixed curvature κ. The model spaces are spheres ($\kappa > 0$), the plane $\mathbb{R}^2$ ($\kappa = 0$) and hyperbolic spaces ($\kappa < 0$), and through comparison with these spaces, we can bound the curvature of the metric space by κ. In this paper we shall use comparison with planar triangles, which will give us curvatures bounded from above by 0.

Given a geodesic metric space X, a *geodesic triangle* in X consists of three points a, b, c and geodesic segments joining them. A planar *comparison* triangle

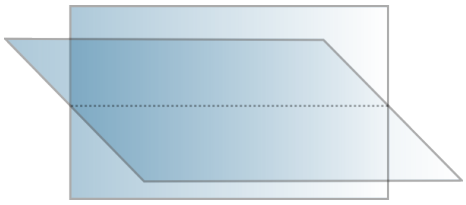


Figure 1.6: Two intersecting planes form a $CAT(0)$ space.

for the triangle abc consists of three points $\bar{a}, \bar{b}, \bar{c}$ in the plane, such that the lengths of the sides in $\bar{a}\bar{b}\bar{c}$ are the same as the lengths of the sides in abc.

A $CAT(0)$ space is a metric space in which geodesic triangles are "thinner" than for their comparison triangles in the plane; that is, $d(x,a) \leq d(\bar{x},\bar{a})$. for any x on the edge bc where $\bar{x}$ is the unique point on the edge $\bar{b}\bar{c}$ such that $d(b,x) = \bar{d}(\bar{b},\bar{x})$ and $d(x,c) = \bar{d}(\bar{x},\bar{c})$. If the planar comparison triangle is replaced by a comparison triangle in the sphere or hyperbolic space of curvature κ, a $CAT(\kappa)$ space is obtained.

A space is said to have non-positive curvature if it is locally $CAT(0)$, i.e. if any point in the space has a radius such that the portion of the space within the radius is $CAT(0)$. Similarly define curvature bounded by κ as being locally $CAT(\kappa)$.

Example 1.8

- The space obtained from intersecting Euclidean spaces as in fig. 1.6 is a $CAT(0)$ space.
- The GPCA construction by Vidal et al. [VMS05] defines a $CAT(0)$ space, giving another potential use of $CAT(0)$ spaces and metric geometry in machine learning.
- The space of phylogenetic trees is a $CAT(0)$ space [BHV01].
- As we are about to see, the space of tree-like shapes is locally a $CAT(0)$ spaces almost everywhere.

One of the main reasons why $CAT(\kappa)$, and $CAT(0)$ in particular, is an attractive property is due to the following result on existence and uniqueness of geodesics.

Proposition 1.9 [BH99, proposition II 1.4] *Let (X,d) be a $CAT(\kappa)$ space. If $\kappa \leq 0$, then all pairs of points have a unique geodesic joining them. For $\kappa > 0$, the same holds for pairs of points at a distance less than $\pi/\sqrt{\kappa}$.*

A more thorough description of the theory of metric geometry can be found in the book by Bridson and Haefliger [BH99].

1.4.2 Curvature in the space of ordered tree-shapes – proof of theorem 1.7

In this section we study the curvature of the shape space using the theory of $CAT(\kappa)$ spaces. We shall show that in a generic set of trees, the shape space will have bounded curvature.

Theorem 1.10 *At a generic point $\bar{x} \in \bar{X}$, the shape space is locally $CAT(0)$, and thus it has locally unique geodesics in a neighborhood of $\bar{x}$.*

Proof. Define a subspace $\bar{W}$ of $\bar{X}$ which is the image under $p\colon X \to \bar{X}$ of the set $W = \{x \in X | \mathrm{pr}_V(x) = 0 \text{ for identified subspaces } V\}$. It is enough to show that the subset of shape space given by $\bar{X} \setminus \bar{W}$ is locally $CAT(0)$.

Pick a point $\bar{x} \in \bar{X} \setminus \bar{W}$. We show that $\bar{X} \setminus \bar{W}$ is locally $CAT(0)$ at $\bar{x}$. Form a second space $\bar{Y}$ by taking, for each identified subspace V with $\bar{x} \cap V \neq \emptyset$, a copy X' of X with a designated subspace V' isomorphic to V, and glue together the X' along the V' as illustrated in fig. 1.6. Give $\bar{Y}$ the quotient pseudometric; this is a metric by using [BH99, Chapter I lemma 5.28] repeatedly, and $\bar{Y}$ is a $CAT(0)$-space, as can be shown by using [BH99, Chapter II theorem 11.3] several times.

Pick $R_{\bar{x}} > r > 0$ such that $B_{\bar{X}}(\bar{x}, r) \cap \bar{W} = \emptyset$. Now $B_{\bar{X}}(\bar{x}, r)$ and $B_{\bar{Y}}(0, r)$ are isomorphic, so in particular, $B_{\bar{X}}(\bar{x}, r)$ is a $CAT(0)$-space. But this proves the first claim.

The second claim, regarding local uniqueness of geodesics, now follows from proposition 1.9. □

Based on this, we are now ready to prove theorem 1.7:

Proof (Proof of theorem 1.7).

i) The QED case: Since $\bar{X}$ is locally $CAT(0)$ at generic points $\bar{x}$, the curvature of $\bar{X}$ is non-positive in a neighborhood U of $\bar{x}$. At points $\bar{x} \in \bar{W}$, however, we will always find pairs of points $\bar{a}_1, \bar{a}_2$ arbitrarily close to x with two geodesics joining them, just as in fig. 1.4c.

ii) The TED case: Consider a tree-shape $\tilde{T} \in \bar{X}$, represented by a point $x \in X$. Induce a second tree-shape $\tilde{T}'$ represented by $x + y_1 + y_2 \in X$, where $y_1, y_2 \in \prod_{e \in E} (\mathbb{R}^d)^n$ such that y_1 and y_2 have one non-zero coordinate, found in different edges, which are both nonzero edges in x. Now, the topology of $\tilde{T}'$ is the same as of $\tilde{T}$. For any $n \in \mathbb{N}$, we can find n TED geodesics $g_1, \ldots, g_n$ from $\tilde{T}$ to $\tilde{T}'$, where g_i can be decomposed as $x \mapsto x + (i/n)y_1 \mapsto x + (i/n)y_1 + y_2 \mapsto x + y_1 + y_2$. It follows that there are infinitely many TED geodesics from $\tilde{T}$ to $\tilde{T}'$. Thus, $(\bar{X}, \bar{d}_1)$ does not have locally unique geodesics anywhere, and as a consequence its curvature is unbounded everywhere [BH99, proposition II 1.4]. □

The practical meaning of theorem 1.7 is that i) we can use techniques from metric geometry to search for QED averages, ii) as we are about to see, for datasets which are sufficiently dense, there exist unique means, centroids and circumcenters for the QED metric, and iii) we cannot use the same techniques in order to prove existence or uniqueness of centerpoints for the TED metric; in fact, any geometric method which requires bounded curvature [Kar77, BH99, BHV01] is going to fail for the TED metric. *This result motivates our study of the QED metric.*

1.4.3 Curvature in the space of unordered tree-shapes

It is easy to prove that the same results also hold for unordered tree-shapes:

Theorem 1.11 *The space $\bar{\bar{X}}$ of unordered trees with the QED metric $(\bar{\bar{X}}, \bar{\bar{d}}_2)$ is non-positively curved. As a consequence, geodesics are locally unique and means, circumcenters and centroids all exist and are unique locally. On the other hand, with the TED metric, $(\bar{\bar{X}}, \bar{\bar{d}}_2)$ has everywhere unbounded curvature, geodesics are not locally unique and the same holds for all the types of average tree discussed in this article.*

1.4.4 Means and related statistics in the space of tree-like shapes

In this section we use what we learned in the previous section to show that, given the QED metric on the space of tree-like shapes, we can find various forms of average shape in the space of ordered tree-like shapes assuming that the data is sufficiently dense.

There are many competing ways of defining central elements given a subset of a metric space. Here discuss several; namely the *circumcenter* considered in [BH99], the *centroid* considered, among other places, in [BHV01], and the *mean* [Kar77].

One way to attack the problems of existence and uniqueness of averages is by using convex functions. Convex functions have minimizers, which are unique for strictly convex functions – hence we can prove existence and uniqueness of averages by expressing them as minimizers of strictly convex functions.

Recall that a function $f\colon [a,b] \to \mathbb{R}$ is convex if $f((1-s)t + st') \leq (1-s)f(t) + sf(t')$ for all $s \in [0,1]$ and $t, t' \in [a,b]$. If we can replace $\leq$ with $<$ whenever $s \in]0,1[$, then f is *strictly* convex, and has a unique minimizer.

We generalize the concepts to geodesic spaces by saying that if X is a geodesic space then a map $f\colon X \to \mathbb{R}$ is (strictly) convex if for any two points $a, b \in X$ and any geodesic $\gamma\colon [0,l] \to X$ from a to b, the function $f \circ \gamma$ is (strictly) convex.

We shall make use of the following standard properties of convex functions:

Lemma 1.12 *i) If $f\colon \mathbb{R} \to \mathbb{R}$ and $g\colon \mathbb{R} \to \mathbb{R}$ are both convex, g is monotonous and increasing, and g is strictly convex, then $g \circ f$ is strictly convex.*

ii) If $f\colon \mathbb{R} \to \mathbb{R}$ and $g\colon \mathbb{R} \to \mathbb{R}$ are both convex, then $g + f\colon \mathbb{R} \to \mathbb{R}$ is also convex. If either f or g is strictly convex, then $g + f$ is strictly convex as well.

The *mean* of a finite subset $\{x_1, \ldots x_s\}$ in a metric space (X, d) is defined as

$$\operatorname{argmin} \sum_{i=1}^{s} d(x, x_s)^2, \tag{1.13}$$

also referred to as the *Frechet mean*. Local minimizers of (1.13) are called *Karcher means*.

Theorem 1.14 *i) If X is a $CAT(0)$-space, then the map $f\colon X \to \mathbb{R}$, $f(x) = d(x,z)$ is convex for any fixed $z \in X$.*

ii) Means exist and are unique in $CAT(0)$-spaces.

Proof. i) Suppose that $x, y \in X$ and let $\gamma\colon [0,l] \to X$ be the unique geodesic from x to y. Form a geodesic triangle with corners at x, y and z; now the edge connecting x and y is parametrized by γ. Form a comparison triangle in the plane with corners x', y' and z'. We have

$$\begin{aligned} f(\gamma(t)) &= f(\gamma((1-t)\cdot 0 + t\cdot 1)) \\ &= d(\gamma(t), z) \\ &\leq d((1-t)x' + ty', z') \\ &\leq (1-t)d(x', z') + td(y', z') \\ &= (1-t)d(x, z) + td(y, z) \\ &= (1-t)f(x) + tf(y), \end{aligned}$$

where the third inequality comes from the fact that if $\gamma(t) = w$ then

$$\begin{aligned} d(w', z') &\leq \min\{d(x', z'), d(y', z')\} \\ &\leq (1-t)d(x', z') + td(y', z'). \end{aligned}$$

ii) Now the function $d_y\colon X \to \mathbb{R}$ given by $d_y(x) = d(x, y)$ is convex for any fixed $y \in X$ by *i)*, and so the function d_y^2 is *strictly* convex by lemma 1.12 *i)*. But then $D = \sum_{i=1}^{s} d_{x_i}^2$ is strictly convex by lemma 1.12 *ii)*, and the mean is just a minimizer of the strictly convex function D, so it exists and is unique. □

We are also going to consider two other types of statistical "prototype" for a dataset, namely *circumcenters* and *centroids*. These are both well-known in the context of metric geometry and $CAT(\kappa)$ spaces.

Definition 1.15 a) **Circumcenters** Consider a metric space (X, d) and a subset $Y \subset X$ with radius r_Y. A point $c \in X$ is a *circumcenter* for Y if $Y \subset \bar{B}(c_Y, r_Y)$, that is, Y is contained in the closed ball of radius r_Y centered at c_Y.

b) **The centroid of a finite set** Suppose that X is a uniquely geodesic metric space. The centroid of a set $S \subset X$ of n elements is defined inductively as a function of the centroids of subsets with $n-1$ elements in the following way:

Suppose $|S| = 2$, and denote the elements of S by s_1 and s_2. The centroid $c(S)$ of S is the midpoint of the geodesic joining s_1 and s_2.

Now suppose $|S| = n$, and denote the elements of S by $s_1, s_2, \ldots, s_n$. Define $c^1(S) = \{c(S') : |S'| = n-1\}$, which is a set with n elements, and similarly, for larger k, $c^k(S) = c^1(c^{k-1}(S))$. All these sets have n elements.

Now if the elements of $c^k(S)$ converge to a point $c \in X$ as $k \to \infty$, then we say that $c = c(S)$ is the *centroid* of S in X.

Based on previous results for $CAT(\kappa)$ spaces, as well, as our results for mean trees, we have for the set of tree-like shapes:

Theorem 1.16 *Endow $\bar{X}$ with the QED metric $\bar{d}_2$. Given a generic point $\bar{x} \in \bar{X}$, there exists a radius $r_{\bar{x}}$ such that sets contained in the ball $B(\bar{x}, r_{\bar{x}})$ have unique means, centroids and circumcenters.*

The same statistical properties also hold for the QED metric on unordered tree-shapes:

At generic points, the space of unordered tree-shapes with the QED metric $(\bar{\bar{X}}, \bar{\bar{d}}_2)$ has local existence and uniqueness of means, circumcenters and centroids. For the TED metric, these are not unique

Proof. Let us start by considering $\bar{X}$ and $\bar{\bar{X}}$ with the QED metric. By Theorems 1.7 and 1.11, $\bar{X}$ and $\bar{\bar{X}}$ are both locally $CAT(0)$ spaces, and by Theorems 1.4 and 1.6 they are both complete metric spaces.

We have just seen that means exist and are unique in $CAT(0)$ spaces, so the statement holds for means.

By [BH99, proposition 2.7], any subset Y of a complete $CAT(\kappa)$ space of radius $r_Y < D_\kappa/2$ has a unique circumcenter, where $D_\kappa = \infty$ whenever $\kappa \leq 0$, and $D_\kappa = \pi/\sqrt{(}\kappa)$ whenever $\kappa > 0$. Hence, the statement holds for circumcenters.

Similarly, by [BHV01, theorem 4.1], finite subsets of $CAT(0)$ spaces X have centroids, which are unique by definition, and the statement also holds for centroids.

Now, let us turn to the TED metric. By definition, for any 2-point dataset, all these definitions of midpoint reduce to finding the midpoint of a geodesic connecting the two points. As we have already seen, geodesics and midpoints are far from unique in the TED metric, which ends the proof. □

1.4.5 Comparison of QED and TED

In this section we discuss the main differences between the TED and QED metrics and compare their performance on the small trees studied in [FLN10].

Geometry

As shown in theorem 1.4 and theorem 1.6 above, both $\bar{X}$ and $\bar{\bar{X}}$ are complete geodesic spaces with both the TED and QED metrics. However, by theorem 1.7, the QED metric gives locally non-positive curvature at generic points, while the TED metric gives unbounded curvature everywhere on $\bar{X}$. Equivalently, geodesics are locally unique almost everywhere in the QED metric, while being nowhere locally unique in the TED metric. As emphasized by theorem 1.16, this means that we cannot imitate the classical statistical procedures on shape spaces using the TED metric, while for the QED metric, we can.

Note, moreover, that the QED metric is the quotient metric induced from the Euclidean metric on the pre-shape space X, making it the natural choice of metric seen from the shape space point of view.

Computation

The TED metric has nice local-to-global properties, as illustrated in fig. 1.7a. If the trees T_1 and T_2 are decomposed into subtrees $T_{1,1}, T_{1,2}$ and $T_{1,2}, T_{2,2}$ as in fig. 1.7a such that the geodesic from T_1 to T_2 restricts to geodesics between $T_{1,1}$ and $T_{2,1}$ as well as $T_{1,2}$ and $T_{2,2}$, then $d(T_1, T_2) = d(T_{1,1}, T_{2,1}) + d(T_{1,2}, T_{2,2})$. This property is used in many TED algorithms, and the same property does not hold for the QED metric, due to the square root involved.

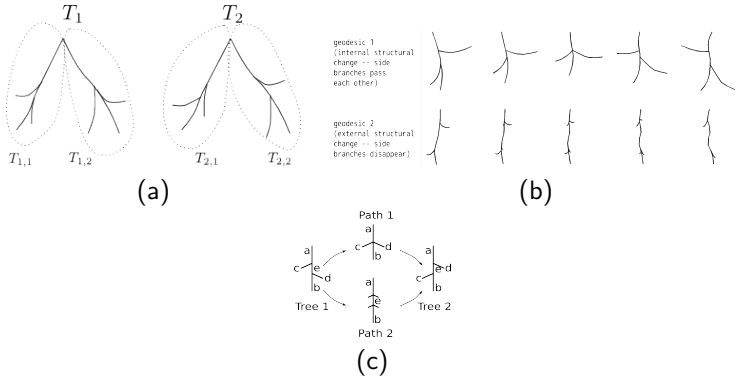


Figure 1.7: (a) Local-to-global properties of TED: $\bar{d}_1(T_1, T_2) = \bar{d}_1(T_{1,1}, T_{2,1}) + \bar{d}_1(T_{1,2}, T_{2,2})$. (b) Generic geodesic deformations: geodesic 1 goes through tree with a trifurcation, while geodesic 2 does not go through internal structural transitions. (c) Two options for structural transition.

Performance

To compare the performance of TED and QED metrics on small, simple trees, consider the two tree-paths in fig. 1.7c, where the edges are endowed with non-negative scalar attributes a, b, c, d, e describing edge length. Path 1 indicates a matching of the identically attributed edges c and d, while Path 2 does not make the match. Now, the cost of Path 1 is $2e$ in both metrics, while the cost of Path 2 is $2\sqrt{c^2 + d^2}$ in the QED metric and $2(c+d)$ in the TED metric. In particular, TED will chose to identify the c and d edges whenever $e^2 \leq c^2 + 2cd + d^2$, while QED makes the match whenever $e^2 \leq \frac{1}{2}(c^2 + d^2)$. That is, TED will be more prone to the internal structural change than QED. This is also seen empirically in the comparison of TED and QED matching in fig. 1.9. Note that although the TED is more prone to matching trees with different tree-topological structures, the matching results are similar, as is also expected since the metrics are quite similar.

1.5 Computation and complexity

Complexity is a problem with computing both TED and QED distances, in particular for $3D$ trees, which do not have a canonical planar order. Here we discuss how to use geometry and anatomy to find approximations of the metric whose complexity is significantly reduced.

1.5.1 Ordered trees: Reducing complexity using geometry

The definition of $\bar{d}$ in (1.1) opens for considering infinitely many possible paths. However, we can significantly limit the search for a geodesic by taking the geometry of tree-space into account.

Theorem 1.17 *i) Tree-shapes that are truly binary (i.e. their internal edges are not collapsed) are generic in the space of all tree-like shapes.*

ii) For generic pairs of trees, the geodesic tree-shape deformations pass through trees which have only bifurcations or trifurcations. Tht is, locally, the geodeics are of the forms shown in in fig. 1.7b.

Proof. i) Let $\tilde{T}$ be a tree-shape in $\bar{X}$ or $\bar{\bar{X}}$ which is not truly binary, which is represented by a binary tree T. By adding arbitrarily small noise to the zero attributes on edges of T, we obtain truly binary tree-shapes $\tilde{T}'$ which are arbitrarily close to $\tilde{T}$. Moreover, the set of full truly binary tree-shapes in $\bar{X}$ or $\bar{\bar{X}}$ is open and dense. Hence, truly binary tree-shapes are generic both in $\bar{X}$ and in $\bar{\bar{X}}$.

ii) Suppose given a pair of trees whose geodesic passes through a tree with a vertex of order 4 or higher. There exists an arbitrarily small deformation of either one of the two endpoint trees which will split the higher order vertex into a series of bifurcations and trifurcations. On the other hand, suppose given a pair of trees whose geodesic only passes through trees with bifurcations and trifurcations. An arbitrarily small deformations of either endpoint will not change the topological structure of the intermediate trees. Thus, intermediate trees in geodesics connecting generic pairs of tree-shapes will only have bifurcations and trifurcations. □

The essense of the theorem is, that *binary tree-shapes are generic, but that does not mean that non-binary trees do not need to be considered! While non-binary trees do not appear as randomly selected trees, they* do *appear in paths between randomly selected pairs of trees*, as in fig. 1.7b. This means that we can work on a less complicated semi-pre-shape-space where we, instead of identifying all possible representations of the same shape, restrict ourselves to making the identifications illustrated by fig. 1.7b.

This is similar to the generic form of shock graphs and generic transitions between shock graphs found by Giblin and Kimia [GK03]. The notions of genericity in the two settings are, however, different since in [GK03], genericity is defined with respect to the Whitney topology on the space of the corresponding shape boundary parametrizations.

Although we do run into non-binary tree-like shapes in real-life applications, for instance when studying airway trees, this can be interpreted as an artifact of resolution rather than as true higher-degree vertices. For instance, airway extraction algorithms record trifurcations when the relative distances are below certain treshold values.

Finally, we often study trees with a somewhat pre-defined overall structure, Thus, it is often safe to assume that the number of internal structural transitions found in a geodesic deformation is low. For the airway trees studied in Section 1.7.2, we find empirically that it is enough to allow for one structural change in each lobar subtree.

Using these arguments, it becomes feasible to computationally handle anatomical structures such as airway trees. In many anatomical applications, such as airways or blood vessels, we are not interested in studying tree-deformations between very large networks or trees, since beyond a certain point, the tree-structure stops following a predetermined pattern and becomes a stochastic variable, where a geodesic analysis does not necessarily make sense.

1.5.2 Unordered trees: Reducing complexity using geometry- or anatomy-based semi-labeling schemes.

It is well known that the general problem of computing TED-distances between unordered trees is NP-complete [ZSS92], and the QED metric is probably generally not less expensive to compute, as indicated also by theorem 1.6. Here we discuss how to use geometry and anatomy to find approximations of the metric whose complexity is significantly reduced.

In particular, *trees appearing in applications are usually not completely unordered, but are often semi-labeled.* Semi-labelings can come from geometric or anatomical properties as in the pulmonary airway trees studied in Example 1.18 below, or may be obtained by a coarser registration method. (Semi-)labelings can also come from a TED distance computation or approximation, which is a reasonable way to detect approximate structural changes since the TED and QED give similar matchings, as seen in fig. 1.9.

Example 1.18 (Semi-labeling of the upper airway tree) In fig. 1.9c, we see a "standard" airway tree with branch labels (or at least the first generations of it). Most airway trees have similar, but not necessarily identical, topological structure to this one, and several branches, especially of low generation, have names and can be identified by experts.

The top generations of the airway tree, shown in red, serve very clear purposes in terms of anatomy. The root edge is the trachea; the second generation edges are the left and right bronchi; and the third generation edges lead to the lung lobes. As these are easily identified, we find a semi-labeling of the airway tree, which is used to simplify the computational complexity for airway distances in Section 1.7.2.

1.6 QED approximation

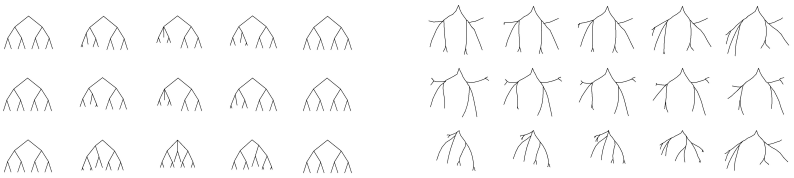
We explain a simple implementation of the QED metric. As shown in the next section, our experimental results are promising.

Alignment of trees and edges.

We translate all edges to start at 0. We do not factor out scale in the tree-space, because in general, we consider scale an important feature of shape. Edge scale in particular is a critical property, as the dynamics of appearing and disappearing edges is directly tied to edge size.

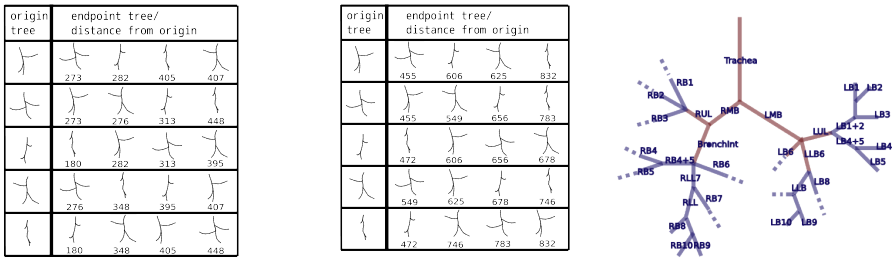
Edge shape comparison.

We represent each edge in a tree by a fixed number of landmark points – in our case 6 – evenly distributed along the edge, the first one at the origin (and hence neglected). The distance between two edge attributes $v_1, v_2 \in (\mathbb{R}^d)^5$ is defined as the Euclidean distance between them. Although simple, this distance measure does take the scale of edges into account to some degree since the edge starting points are aligned.



(a) Topological illustration. (b) Tree-shape illustration.

Figure 1.8: Tree-space paths with three different options for structural changes in the left hand side of the depth 3 tree.



(a) Matching in the QED metric. (b) Matching in the TED metric. (c)

Figure 1.9: (a-b) Given a set of five data trees, we match each to the four others in both metrics. (c) Standard airway tree structure with branch labels.

Implementation.

We implement the metric for ordered depth 3 trees using Algorithm 1, where the number of internal topological transitions taking place throughout the geodesic deformations is limited to either 1 or 2. This leaves us with the options for structural changes illustrated in figs. 1.8a and 1.8b, applied to the left and right half tree. With respect to the definition of the metric in equation 1.1, this corresponds to bounding the number k, giving:

$$\bar{d}(\bar{x},\bar{y}) = \inf_{k\leq K}\left\{\sum_{i=1}^{k} d(x_i,y_i)|x_1\in\bar{x}, y_i\sim x_{i+1}, y_k\in\bar{y}\right\}, \tag{1.19}$$

where we have used $K = 2$ or 3. Geometrically, K is the number of Euclidean segments concatenated to form the geodesic, which equals $1 +$ *the number of internal topological transitions made throughout the geodesic.*

The complexity of Algorithm 1 is $O(n^{K-1}\cdot C(n,K))$, where n is the number of internal vertices, K is the bound on k and $C(n,K)$ is the maximal complexity of the optimization in line 8. For the unordered airway trees in section 1.7.2, we combine Algorithm 1 with a complete search through the set of branch orderings.

Algorithm 1 Computing QED approximations between ordered, rooted trees with up to $k = K - 1$ structural transitions

x, y planar rooted depth n binary trees
$\mathbf{S} = \{S\}$ set of ordered identified pairs $S = \{S_1, S_2\}$ of subspaces of X corresponding to internal topological changes, corresponding to a subspace S of $\bar{X}$, s.t. if $\{S_1, S_2\} \in \mathbf{S}$, then also $\{S_2, S_1\} \in \mathbf{S}$.
for $\tilde{S} = \{S^1, \ldots, S^s\} \subset \mathbf{S}$ with $|\tilde{S}| \leq k$ **do**
  for $p^i \in S^i$ with representatives $p_1^i \in^i S_1$ and $p_2^i \in S_2^i$ **do**
    $p = (p^1, p^2, \ldots, p^s)$
    $f(p) = \min\{d_2(x, p_1) + \sum_{j=1}^{s-1} d_s(p_i^2, p_{i+1}^i) + d_2(p_s, y)\}$
  end for
  $d_{\tilde{S}} = \min \left\{ f(p) \;\middle|\; \begin{array}{c} p = (p^1, \ldots, p^s), p^i \in S^i, \\ \tilde{S} = \{S^1, \ldots, S^s\} \end{array} \right\}$
  $p_{\tilde{S}} = \{p_1^i, p_2^i\}_{i=1}^s = \text{argmin} f(p)$
end for
$d = \min\{d_S | \tilde{S} \subset \mathbf{S}, |\tilde{S}| \leq k\}$
$p = \{p_1, p_2\}_{i=1}^s = \{p_{\tilde{S}} | d_{\tilde{S}} = d\}$
geodesic $= g = \{x \to p_1^1 \sim p_2^1 \to p_1^2 \sim p_2^2 \to \ldots \to p_1^s \sim p_2^s \to y\}$
return d, g

1.7 Experimental results

The QED metric is new, whereas the matching properties of the TED metric are well known [SKK04]. In this section we present experimental results on real and synthetic data which illustrate the geometric properties of the QED metric. The experiments on airway trees in Section 1.7.2 show, in particular, that it is feasible to compute the metric distances between real, $3D$ data trees.

1.7.1 Synthetic planar trees of depth 3

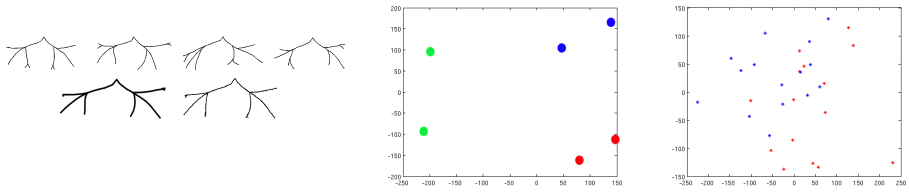
In order to illustrate the geometrically inutitive behaviour of the geodesic deformations, and their ability to handle internal topological differences, we have uploaded movies illustrating geodesics between planar depth 3 trees, as well as a matching table for a set of planar depth 3 trees, to the webpage

http://image.diku.dk/aasa/tree_shape/planar.html.

Moreover, we have computed some simple mean and centroid trees, shown in fig. 1.10a.

1.7.2 Results in 3D: Pulmonary airway trees

We also compute QED distances between subtrees of pulmonary airway trees. The airway trees were first segmented from low dose screening computed tomography (CT) scans of the chest using a voxel classification based airway tree segmentation algorithm by Lo et al. [LSA+10]. The centerlines were extracted from the segmented airway trees using a modified fast marching algorithm based on [SLC+02]. The method gives a tree structure directly through connectivity of parent and children branches. Leaves with a volume less than 11 mm^3 were



(a) (b) (c)

Figure 1.10: (a) Top: A small set of synthetic planar trees. Bottom left: The mean tree. Bottom right: The centroid tree.(b) The separation of the three patients is clearly visible. (c) Although the classes overlap, there is a clear tendency to separation between healthy smokers and COPD patients.

assumed to be noise and pruned away, and the centerlines were sampled with 6 landmark points on each edge.

First experiment: Separation of different patients

Six airway trees extracted from CT scans taken from three different patients at two different times were analyzed, restricted to the first six generations of the airway tree. The first three generations were identified and labeled as in Ex. 1.18, leaving us with topological variation in four lobar depth 3 subtrees representing the 4^{th} to 6^{th} airway tree generations. Algorithm 1 was run with both $K = 2$ and $K = 3$, giving the same results for all trees. Thus, we find it perfectly acceptable to restrict our search to paths with low values of K.

Based on the inter-airway tree distances we made the 2-dimensional multi-dimensional scaling plot shown in fig. 1.10b, which clearly shows a separation of the three different patients.

Second experiment: Tree-shape and disease

It is well-known that there is correlation between geometric airway measures such as airway wall thickness and COPD [WDE+09], and in a recent study such measures are successfully combined with global anatomical knowledge of the airway geometry [SLD+11]. We perform a different experiment to investigate correlation between COPD and the airway tree-shape based only on QED distance, i.e. without incorporating the airway wall thickness.

We compute coarse approximate distances (with $K = 2$) between 30 airway trees, 15 from healthy male smokers and 15 from male smokers with moderate COPD and no emphysema. Based on the resulting distances, we obtain the multidimensional scaling plot found in fig. 1.10c, which shows a clear tendency in the data even though the two classes seem to overlap.

1.8 Conclusions and future work

Starting from a purely geometric point of view, we define a shape space for tree-like shapes. The intuitive geometric framework allows us to simultaneously take both global tree-topological structure and local edgewise geometry into

account. We define two metrics on this shape space, QED and TED, which give the shape space a geodesic structure (theorems 1.4 and 1.6).

QED is a geometrically natural metric, which turns out to have geometric properties which are essential for statistical shape analysis. In particular, the QED metric has local uniqueness of geodesics and local existence and uniqueness for three versions of average shape, namely the mean, the circumcentre and the centroid (theorem 1.16). TED does not share these properties, but has somewhat better computational properties.

Both metrics are generally NP hard to compute for $3D$ trees. We explain how semi-labeling schemes and geometry can be used to overcome the complexity problems, and illustrate this by computing QED distances between trees extracted from pulmonary airway trees as well as synthetic planar data trees.

Our future research will be centered around two points: Development of nonlinear statistical methods for the singular tree-shape spaces, and finding fast approximations and heuristics for the QED metric. The latter will pave the road for computing averages and modes of variation for large, real $3D$ data trees – which is by no means trivial, due to the complexity of computing exact distances.

1.9 Appendix: Proof of theorem 1.4

1.9.1 Precise shape space definition

The general idea of the construction of the tree-shape space is illustrated in the section preceding theorem 1.4. However, in order to write out the technicalities of the proof, we need to be a little more specific about how the identifications are taking place:

For each combinatorial type of tree-shape C_j $(j = 1, \ldots, K)$ which can be represented using $\mathscr{T}$, there is a family E_j^i of subsets of E, which satisfy

a) $\sharp E_j^1 = E_j^i$ for all $i = 1, \ldots, n_j$, $j = 1, \ldots, K$.

b) there is a depth-first order on each E_j^i induced by the depth-first order on E, such that the ordered, combinatorial structure defined by any E_j^i coincides with that defined by E_j^1.

The subset E_j^i for any i lists the set of edges in $\mathscr{T}$ which have nonzero attributes for the i^{th} representation of any shape of type C_j. Corresponding to each E_j^i is the linear subspace Z_j^i of X given by

$$Z_j^i = \{(a_e) \in \prod_{e \in E} (\mathbb{R}^d)^n | a_e = 0 \text{ if } e \notin E_j^i\}$$

and by condition b) we can define isometries $\phi_j^i \colon Z_j^i \to \prod_{e \in E_j^1} A$ by forgetting the zero entries in Z_j^i and keeping the depth-first coordinate order. We generate the equivalence $\sim$ on X by asking that $z \sim w$ whenever $\phi_j^i(z) = \phi_j^l(w)$ for some i, j, l. We now define the space of ordered tree-like shapes as the quotient $\bar{X} = X/\sim$, and define the quotient map $p \colon X \to \bar{X}$.

1.9.2 The pseudometric is a metric

To make the proof as clear and readable as possible, it has been split into several propositions and lemmas below, and is summarized after all the smaller results have been proven.

Proposition 1.20 *The pseudometric $\bar{d}$ is a metric on $\bar{X}$.*

Proof. Since $\bar{d}_1$ and $\bar{d}_2$ are pseudometrics, which is easily seen from the definition, it suffices to show that $\bar{d}_i(\bar{x}, \bar{y}) = 0$ implies $\bar{x} = \bar{y}$. Moreover, it is also easy to show that $d_1(\bar{x}, \bar{y}) \geq d_2(\bar{x}, \bar{y})$ for any $\bar{x}, \bar{y} \in \bar{X}$, so it suffices to show that $\bar{d}_2(\bar{x}, \bar{y}) = 0$ implies $\bar{x} = \bar{y}$.

Hence, from now on, write $\bar{d}$ for $\bar{d}_2$, and assume that $\bar{d}(\bar{x}, \bar{y}) = 0$.

Choose $\epsilon > 0$ such that

$$\epsilon \ll \min \left\{ \begin{array}{c|c} \begin{array}{c} \|x_e\| > 0, \\ \|y_e\| > 0, \\ \|x_e - x_{\tilde{e}}\| > 0, \\ \|y_e - e_{\tilde{e}}\| > 0, \end{array} & \begin{array}{c} x = (x_e) \in \bar{x}, \\ y = (y_e) \in \bar{y} \end{array} \end{array} \right\}. \tag{1.21}$$

We may assume that $x, y \in \bigcup_{i,j} Z_j^i$ since otherwise we may assume by symmetry that $\bar{x} = \{x\}$ and $\bar{d}(\bar{x}, \bar{y}) \geq \min\{d(x, \bar{y}), d(x, \bigcup_{i,j} Z_j^i)\} > 0$.

Denote by $\bar{X}_j$ the image of Z_j^i under the quotient projection $p\colon X \to \bar{X}$ for any i.

We may assume that there exists i, j such that

$$\bar{x} \cap Z_j^i \neq \emptyset \neq \bar{y} \cap Z_j^i \tag{1.22}$$

since otherwise,

$$\bar{y} \cap \left(\bigcup \{ Z_j^i | \bar{x} \cap Z_i^j \neq \emptyset \} \right) = \emptyset.$$

Now, in that case, $\bar{y}$ is a finite set, and $\bigcup \{ Z_j^i | \bar{x} \cap Z_j^i \neq \emptyset \}$ is a closed set, so

$$d\left(\bar{y}, \bigcup \{ Z_j^i | \bar{x} \cap Z_j^i \neq \emptyset \} \right) > 0.$$

The path will have to go through some $Z_{\tilde{j}}^{\tilde{i}}$ which does not contain points equivalent to y, and

$$d\left(\bar{y}, \bigcup \{ Z_j^i | \bar{y} \cap Z_j^i = \emptyset \} \right) > \epsilon,$$

because in order to reach $\bigcup \{ Z_j^i | \bar{y} \cap Z_j^i = \emptyset \}$ we need to remove some edge attributes which are nonzero in $\bar{y}$, and $\epsilon \ll \|y_e\|$ for all $y_e \neq 0$.

But if the "path points" stay in $\bar{X}_j$, then the path consists of shifting and changing the nonzero edge attributes of the trees in question, and the only way that that will give a sum $< \epsilon$ is if the trees are identical and the path is constant. □

1.9.3 Topology of the space of tree-like shapes

In this section we prove the following proposition, which is part of theorem 1.4:

Proposition 1.23 *The tree-shape space $(\bar{X}, \bar{d})$ is a complete proper geodesic space, and $\bar{X}$ is contractible.*

Before turning to the proof, we note that although $\bar{X}$ is no longer a vector space like X, there is a well-defined notion of size for elements of $\bar{X}$, stemming from the norm on X:

Lemma 1.24 *Note that if $x \sim y$, we must have $\|x\| = \|y\|$; hence we can define $\|\bar{x}\| := \|x\|$.*

Proof. The equivalence is generated recursively from the conditions $x \sim y$ whenever either $x = y$, indicating $\|x\| = \|y\|$; or $\phi^i_j(x) = \phi^i_k(y)$, indicating $\|x\| = \|\phi^i_j(x)\| = \|\phi^i_k(y)\| = \|y\|$ since the ϕ are isometries. Hence, the lemma holds by recursion. □

We now turn to the proof of proposition 1.23.

Proof. We shall prove that $(\bar{X}, \bar{d})$ is a proper geodesic space using the Hopf-Rinow theorem for metric spaces:

Theorem 1.25 [BH99, Chapter I, proposition 3.7] *Every complete locally compact length space is a proper geodesic space.*

A length space is a metric space in which the distance between two points can always be realized as the infimum of lengths of paths joining the two points. Note that this is a weaker property than being a geodesic space, as the geodesic joining two points does not have to exist; it is enough to have paths that are arbitrarily close to being a geodesic. It follows from [BH99, Chapter I lemma 5.20] that $(\bar{X}, \bar{d})$ is a length space for *any* metric d on X.

To see that the tree-shape space is locally compact, note that the projection $p\colon X \to \bar{X}$ is finite-to-one, so any open subset U of $\bar{X}$ has as preimage a finite union $\cup_{i=1}^N U_i$ of open subsets of X, such that $\operatorname{diam}(U_i) = \operatorname{diam}(U)$ and $p(\cup_i \bar{U}_i) = \bar{U}$ is compact whenever U is bounded.

We also need to prove that $(\bar{X}, \bar{d})$ is complete:

Proposition 1.26 *Let $\bar{d}$ denote either of the metrics $\bar{d}_1$ and $\bar{d}_2$. The shape space $(\bar{X}, \bar{d})$ is complete.*

In the proof of proposition 1.26 as well as later completeness proofs we shall use the following lemma from general topology:

Lemma 1.27 [Dug66, Chapter XIV, theorem 2.3] *Let (X, d) be a metric space and assume that the metric d has the following property:*

There exists $\epsilon > 0$ such that for all $y \in Y$ the closed ball $\bar{B}(y, \epsilon)$ is compact. Then $\bar{d}$ is complete.

Lemma 1.28 *Bounded closed subets of $\bar{X}$ are compact.*

Proof. By assumption X contains the origin 0 in $\prod_{e \in E} B$; hence, any closed, bounded subspace C in $\bar{X}$ is contained in the ball $\bar{B}_{\bar{d}}(\bar{0}, R)$ in $\bar{X}$ for some $R > 0$, where $\bar{0}$ is the image $p(0) \in \bar{X}$. Now, since $\|x\| = \|\bar{x}\|$, it follows that $p^{-1}(\bar{B}_{\bar{d}}(\bar{0}, R)) = \bar{B}_d(0, R)$, which is closed and bounded. Now, since closed, bounded subsets of A are compact, also closed, bounded subsets of $X = \prod_{e \in E} A$ must be compact, so $\bar{B}_d(0, R)$ is compact, and by continuity of p, $\bar{B}_{\bar{d}}(\bar{0}, R))$ is compact. But then C is compact also. □

But then it is very easy to prove proposition 1.26:

Proof (Proof of theorem 1.26). By lemma 1.28, all closed and bounded subsets of $\bar{X}$ are compact, but then by lemma 1.27 the metric $\bar{d}$ must be complete.□

Use the Hopf-Rinow theorem as stated in theorem 1.25 we thus prove that $(\bar{X}, \bar{d})$ is a complete, proper geodesic space. The contractibility follows from the following lemma:

Lemma 1.29 *Let B be a normed vector space and let $\sim$ be an equivalence on B such that $a \sim b$ implies $t \cdot a \sim t \cdot b$ for all $t \in \mathbb{R}$. Then $\bar{B} = B/\sim$ is contractible.*

Proof. Define a map $H \colon \bar{B} \times [0,1] \to \bar{B}$ by setting $H(\bar{x}, t) = t \cdot \bar{x}$. Now H is well defined because of the condition on $\sim$, and $H(\bar{x}, 0) = 0 \ \forall \ \bar{x} \in \bar{B}$ so H is a homotopy from $\mathrm{id}_{\bar{B}}$ to the constant zero map. □

With this, we conclude the proof of proposition 1.23. □

Combining the theorems of Section 1.3.3 together, we see that the proof of theorem 1.4 is complete.

Chapter 2

Computing averages in spaces of tree-like shapes

The mean is often the most important statistic of a dataset as it provides a single point that summarizes the entire set. While the mean is readily defined and computed in Euclidean spaces, no commonly accepted solutions are currently available in more complicated spaces, such as spaces of tree-structured data. In this paper we study the notion of means, both generally in Gromov's $CAT(0)$-spaces (metric spaces of non-positive curvature), but also specifically in the space of tree-like shapes. We prove local existence and uniqueness of means in such spaces and discuss three different algorithms for computing means.

We make an experimental evaluation of the three algorithms through experiments on three different sets of data with tree-like structure: a synthetic dataset, a leaf morphology dataset from images, and a set of human airway subtrees from medical CT scans. This experimental study provides great insight into the behavior of the different methods and how they relate to each other. More importantly, it also provides mathematically well-founded, tractable and robust "average trees". This statistic is of utmost importance due to the ever-presence of tree-like structures in human anatomy, e.g., airways and vascularization systems.

2.1 Notions of means

Centroids, weighted averages, midpoints of a pair of points, and other variations on the sample mean are the basic building blocks of statistical computations. While they are simple to compute when the underlying sample space is Euclidean, they may become much more complex in non-linear sample spaces. A classical definition of centroids in Euclidean space, dating back to Appolonios of Perga, has a direct extension to general metric spaces [Kar77, JO10]: a mean of the finite collection $(x_i)_i$ of points in a metric space (X, d) is a minimizer of the function

$$\Phi(x) = \sum_{i=1}^{n} d(x, x_i)^2. \tag{2.1}$$

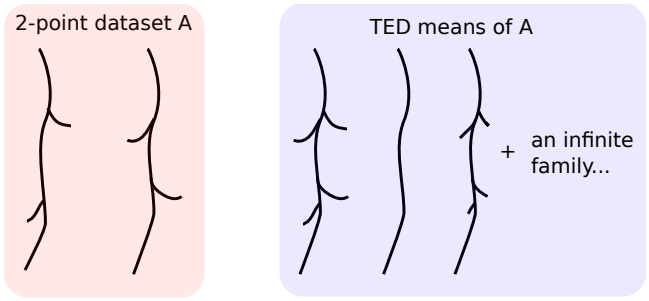


Figure 2.1: The infinite family of trees on the right are TED means for the set of two trees on the left.

A local minimizer of Φ is called a *Karcher mean* while a global is called a *Fréchet mean*. But when does such a minimizer exist? When is it unique? Although the above definition does not require existence of geodesics, this is often needed in order to compute a minimizer. This reveals important problems already in the simplest of situations. If geodesics exist in X, a solution to the above problem for a set of two points a and b is the point c on the geodesic segment from a to b such that $d(a, c) = d(b, c)$. But what if there is more than one geodesic segment between a and b? The midpoint of each geodesic segment will minimize eq. 2.1.

A key example where this problem occurs is the (Tree) Edit Distance (denoted TED in the sequel) used in spaces of attributed graphs and trees, e.g., shock graphs [KTSK00, FVS$^+$10, TK10]. This metric is problematic as *even locally*, geodesics (edit paths) are not unique, and this prevents the existence of well-defined means even in a local context. For the pair of trees on the left in fig. 2.1 there is an infinite family of geodesics (and hence means) generated by varying the order and amount by which the side branches are shrunk and grown while deforming one tree into the other. A common approach for choosing a typical representative using TED is to choose the simplest possible mean, in this case the one shown in the middle. When iteratively computing means, however, one risks ending up with mean trees that are significantly simpler than the trees in the dataset. This explains the reduced complexity of the TED means found by Trinh and Kimia [TK10]. Similarly, in the graph embedding work of Bunke and collaborators [FVS$^+$10, RB09b], severe restrictions on the solution space and geodesics (edit paths) have to be imposed in order to have well defined constructions as simple as midpoints; restrictions that are kept implicit.

Even when geodesics are locally unique, means need not be unique. Consider data uniformly distributed along the equator of a sphere: both poles will minimize eq. 2.1. In shape analysis, this problem often occurs as most shape spaces are non-linear. In Riemannian manifolds, such as the sphere example, existence of at least locally minimizing geodesics is guaranteed. Tools from differential calculus are available for the optimization of eq. 2.1 and Karcher [Kar77] provides conditions for local existence and uniqueness of the mean. Hence, in most statistical contexts one is content to find a (reasonably large) dataset radius within which geodesics and local minimizers of eq. 2.1 exist and are unique; for the unit sphere this radius is $\pi/2$.

The Appolonios problem (eq. 2.1) in Euclidean space can also be solved by geometric constructions using only geodesics and weighted midpoints, which carry over to more general metric spaces. But do different methods provide the same solutions? Even locally? Already in the Riemannian framework, non-linearity introduces difficulties in other constructions. A simple example is Principal Component Analysis, which in the Euclidean case can equally well be defined via maximization of projected variances on subspaces *or* minimization of reconstruction errors. But *these approaches lead to different solutions* in the Riemannian setting, e.g., Fletcher's PGA [FLPJ04] and Huckemann's GPCA [HHM10] are not equivalent. With this in mind, one should tread carefully when generalizing from Euclidean methods.

Tree-like shapes (and more generally graph-like shapes), among others, present a great challenge, as they are not naturally modeled as elements of smooth manifolds. Any tree-shape can be obtained as a limit of a large number of tree-shapes with very different tree topologies, creating natural self-intersections in tree-space. This may prevent the use of smooth optimization methods. However, this complexity does not prevent the use of treelike shapes. In computer vision they appear in skeletons and shock-graphs for 2D shape recognition and classification [DSD09, KTSK00, TK10], and they are often encountered in medical imaging, as airways and blood vasculatures have natural tree-like shapes [MKS$^+$09, TMP$^+$05].

Feragen et al. [FLL$^+$10] introduced a construction of tree-like shape spaces with a metric called *Quotient Euclidean Distance* (QED), which gives existence and local uniqueness of geodesics. This follows from the fact that they are locally *CAT(0)-spaces* or *spaces of non-positive curvature*, a concept introduced by Gromov [Gro87] and discussed in the monograph [BH99]. Billera, Holmes and Vogtmann [BHV01] have proposed a $CAT(0)$-space structure for phylogenetic trees, but these trees are abstract objects not encoding 2D or 3D shape, with much more restricted variations.

Spaces of tree-like shapes, in spite of their complexity, offer a very good framework for computing means through geometric solutions to the Appolonios problem. In this paper we explore three such constructions: the centroid, Birkhoff shortening and weighted midpoints. The methods are tested on leaf vasculature shapes and airway tree shapes.

The rest of the article is organized as follows. In sec. 2.2 we discuss the space of tree-like shapes along with a metric that gives locally unique geodesics. We review the basics of metric geometry in spaces of non-positive curvature in sec. 2.3. This leads us to the theoretical novelty of the paper as we prove that unique means exist in such spaces in sec. 2.3.1. We then discuss various ways of defining and computing means (sec. 2.4) followed by an experimental comparison of these means (sec. 2.5). Finally, the paper is concluded with a brief summary and a discussion of open problems.

2.2 The space of tree-like shapes

We are interested in spaces of tree-like shapes as defined by Feragen et al. [FLL$^+$10]. Tree-like shapes are represented as rooted, ordered, binary trees $T = (V, E, r, <)$ with edge attributes $f\colon E \to \mathbb{R}^n$. The attributes take values in $\mathbb{R}^n$, and describe the shape of the particular edge, e.g., via landmark points. The branch order

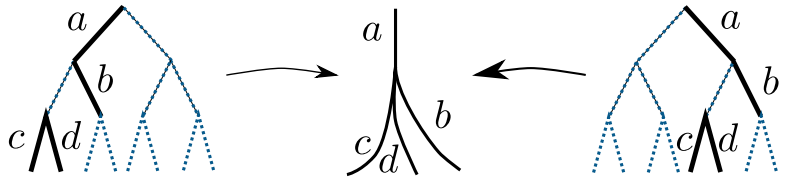


Figure 2.2: Higher-order vertices can be represented by the binary tree by collapsing internal branches, as the dotted blue lines.

can, for instance, come from a planar ordering of branches. To compare the tree-shapes within one single shape-space, all tree-shapes are parametrized using the same combinatorial tree T, which is *sufficiently large* to describe all the tree-shapes in our dataset. Trees with fewer edges are represented by collapsing redundant branches, as in fig. 2.2. Higher-order bifurcations are represented in a similar fashion, also using collapsed branches. This gives a representation space $X = \prod_{e \in E} \mathbb{R}^n$, where all tree-like shapes are represented at least once. A space of ordered tree-like shapes $\bar{X}$ is defined as a quotient space of X by identifying different order-preserving representations of the same tree-shape. This corresponds to folding the Euclidean representation space and gluing it along different representations of the same trees. From the Euclidean metric on X, Feragen et al. induce the *quotient metric* on $\bar{X}$, which in this case is called the *QED metric*. The quotient metric is a standard mathematical construction [BH99], which here creates a piecewise Euclidean metric. If the Euclidean metric on X is replaced with an l_1 product metric, the TED metric is retrieved as quotient metric on $\bar{X}$.

Planar trees come with a given branch order (left to right) on the set of children of each branch, and hence a branch order on the entire tree is easily obtained. Trees that reside in 3D space are not ordered in the same way. In order to optimally compute the distance between two trees, it is necessary to consider all possible orders on the two trees. At a first glance, this seems to give problems with computational complexity. However, when studying shapes that are close together this is not necessary. We can induce an order on each tree by fixing an order on one (sufficiently large) tree, aligning the other trees with the chosen one. The alignment can be done by finding the order that minimizes the distance between the trees. This way, we effectively reduce our set of unordered trees to a set of ordered trees. This becomes a great computational relief when we start computing average $3D$ trees, since we make a large number of distance computations as part of the iterative midpoint procedures.

2.3 Curvature and means in metric spaces

A very favorable property of the space of tree-like shapes is that its local geometry facilitates statistical computations. To be more precise, at "generic points" (that is, at any randomly chosen point) the space is locally $CAT(0)$. This concept is at the heart of our analysis, and is novel in the context of computer vision. Thus, we shall dedicate a few lines to explaining what $CAT(0)$ spaces

are and why they are nice geometric objects.

One way of studying the geometry of general metric spaces is to compare them to spaces whose geometry we understand well, referred to as *model spaces*. The model spaces are spheres (positively curved), the Euclidean plane (flat, no curvature) and hyperbolic spaces (negatively curved). Since metric spaces can be rather pathological, a typical approach to defining curvature is to bound the curvature of the space at a given point from above or below. In this article we study spaces of non-positive curvature, i.e., their curvature is bounded from above by 0. Due to the curvature bounded by 0, we study the metric spaces by comparing triangles in the metric space with triangles in the Euclidean plane, as in Gromov's metric geometry [Gro87, BH99]. Mathematically, this is expressed by the $CAT(0)$ *condition*:

Definition 2.2 ($CAT(0)$ spaces, non-positive curvature) Let X be a geodesic metric space, that is, a space in which all points can be joined by a geodesic. The length of a path is defined by the metric d in X. A *geodesic triangle* abc in X consists of three points a, b and c in X, along with geodesic paths joining the points: $[ab]$, $[bc]$ and $[ac]$, see fig. 2.3. There exists a triangle $\bar{a}\bar{b}\bar{c}$ in the Euclidean plane with vertices $\bar{a}$, $\bar{b}$ and $\bar{c}$ and with edges $[\bar{a}\bar{b}]$, $[\bar{b}\bar{c}]$ and $[\bar{a}\bar{c}]$, whose lengths are the same as the lengths of $[ab]$, $[bc]$ and $[ac]$. This is a *comparison triangle* for abc, see fig. 2.3.

For any point x sitting on the segment $[bc]$, there is a corresponding point $\bar{x}$ on the segment $[\bar{b}\bar{c}]$ in the comparison triangle, such that $\|\bar{x} - \bar{b}\| = d(x, b)$. If we have

$$d(x, a) \leq \|x - a\| \tag{2.3}$$

for every such x, then the geodesic triangle abc satisfies the $CAT(0)$ *condition*.

The metric space X is a $CAT(0)$ space if any geodesic triangle abc in X satisfies the $CAT(0)$ condition given in eq. 2.3. Geometrically, this means that triangles in X are *thinner* than triangles in $\mathbb{R}^2$. Spaces which are locally $CAT(0)$ are ***non-positively curved.***

A few examples of $CAT(0)$ spaces are:

1) Euclidean space is a $CAT(0)$ space.

2) Any non-positively curved manifold is locally $CAT(0)$.

3) The union of two intersecting planes is a $CAT(0)$ space [BH99, Chapter II Theorem 11.3].

4) The space of tree-like shapes, see fig. 2.4, is locally $CAT(0)$ at generic points [FLL$^+$10, Theorem 2].

2.3.1 Means in $CAT(0)$ spaces

As shown in [FLL$^+$10] and discussed above, for generic points in the space of tree-like shapes there exists a radius $r_x > 0$ such that the ball $B(x, r_x)$ is a $CAT(0)$-space. For a point x in the shape space in fig. 2.4, whose distance to the projected origin is d, $r_x = d\tan(\pi/8)$; namely the radius within which all points are joined by a unique geodesic. Hence, if means exist and are unique in $CAT(0)$ spaces, they exist and are unique for sufficiently dense sets of tree-like shapes.

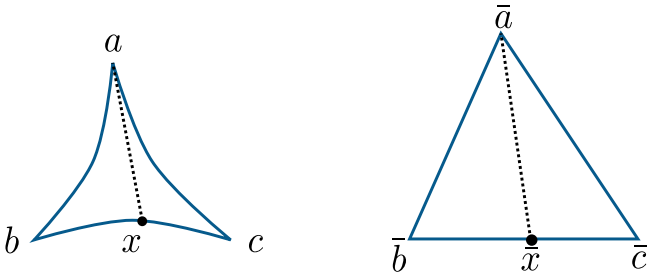


Figure 2.3: Left: A geodesic triangle, right: the corresponding comparison triangle in the plane $\mathbb{R}^2$.

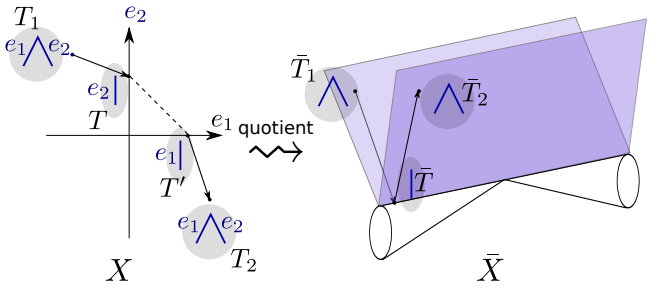


Figure 2.4: The simplest nontrivial example of a tree-shape space $\bar{X}$ (right) with representation space X (left): Trees with two edges e_1 and e_2, with scalar edge attributes. Here we see the path from $\bar{T}_1$, with two edges, through $\bar{T}$, with one edge, to $\bar{T}_2$, also with two edges. The one-branch tree $\bar{T}$ has two representatives in X on the e_1 and e_2 axes, respectively, which are glued together in the shape space (right). Note how the path from $\bar{T}_1$ to $\bar{T}_2$ in $\bar{X}$ is made from concatenated Euclidean lines in X, with a "teleportation" (dotted line) gluing the representatives T and T' of $\bar{T}$ together.

Theorem 2.4 *Means exist and are unique in $CAT(0)$-spaces.*

Proof: Let (X, d) be a $CAT(0)$-space. It follows from the $CAT(0)$-inequality that given any fixed $y \in X$, the function $d_y \colon X \to \mathbb{R}$ given by $d_y(x) = d(x, y)$ is convex. Then the function d_y^2 is also convex since the function $g \colon a \mapsto a^2$ is monotone, increasing and convex. Moreover, $((1-s)d_y(x) + sd_y(x'))^2 < (1-s)d_y^2(x) + sd_y^2(x')$ for any $s \in]0, 1[$, and hence, d_y^2 is actually strictly convex. Then $\Phi = \sum_{i=1}^{s} d_{x_i}^2$ is strictly convex, and the mean is a minimizer of the strictly convex function Φ; hence it exists and is unique. □

As a direct consequence, *means do exist and are even unique for datasets of sufficiently small diameter in tree-space with the QED metric.* This is in stark contrast to the TED metric where means are practically never unique.

2.4 Computing means

For finite point sets in the Euclidean space, the easiest way to compute the mean is the closed form solution $\sum x_i/N$. This solution does not carry over to non-linear spaces, but other techniques do.

One way to optimize eq. 2.1 is to use gradient descent. However, in the case of tree-structured data, the shape space is not even a smooth manifold. Hence, we cannot perform a regular gradient descent, but would have to develop optimization methods for non-smooth spaces. Moreover, we do not have an analytic expression for the gradient even in the smooth parts of the space. In order to numerically approximate derivatives, the distance function would need to be evaluated a large number of times. This makes optimization schemes unattractive for computing means in tree-space, where distances are expensive to compute.

Since the two most obvious methods for computing a mean shape are not applicable in tree-space, we need to look for alternative ways of computing means, which do not require evaluating too many distances. In this paper we study three iterative algorithms, which are all based on halving geodesics, namely centroids, Birkhoff shortening and weighted midpoints.

2.4.1 The centroid

The *centroid* $c(A)$ of a dataset $A = \{x_1, \ldots, x_N\}$ in a geodesic metric space (X, d), is defined as follows by Billera, Holmes and Vogtmann [BHV01]:

If $N = 2$, then $c(A)$ is the midpoint of the geodesic connecting the two points x_1 and x_2. Assume that we have a working definition of centroid for datasets with at most $N-1$ points. Then define a set of subsets $A_1, \ldots, A_N$ by setting $A_i = A \setminus \{x_i\}$. Define a new N-element set $c^1(A) = \{c(A_1), \ldots, c(A_N)\}$, that is replacing A by centroids for each $(N-1)$-element subset of A. This is illustrated in fig. 2.5. Define a sequence of sets $c^k(A)$ for $k \in \mathbb{N}$ by setting $c^k(A) = c^1(c^{k-1}(A))$; if the sequence $c^k(A)$ converges to a single point $c \in X$, then c is the *centroid* of A.

It is easy to see that in Euclidean space, the centroid is just the regular mean $\sum_i x_i/N$. Billera, Holmes and Vogtmann [BHV01, Theorem 4.1] prove that in $CAT(0)$-spaces, the centroid construction converges to a unique point.

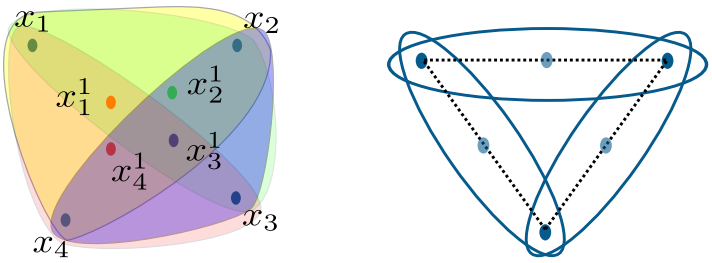


Figure 2.5: This illustration is best viewed in color. The centroid process is defined recursively with respect to the number of elements in the dataset. On the 4 point dataset $A = \{x_1, x_2, x_3, x_4\}$, each iteration consists finding the centroids of four 3-point subsets of A. Left: The iterative process illustrated for a 4 point dataset, and the centroids are denoted $x_1^1, x_2^1, x_3^1, x_4^1$. Right: For datasets with 3 points, the centroid process coincides with Birkhoff shortening.

It is, however, unknown whether this point is generally the mean, as defined in eq. 2.1, or a different point.

Algorithmic properties

The centroid is nice in theory since it defines a well-posed problem: centroids exist and are unique in $CAT(0)$ spaces. However, its algorithmic properties are not attractive. The computational complexity of computing $c(A)$ for a dataset A with N elements, is of the order N times the computational complexity of computing $c(A')$ for a dataset A' with $N-1$ elements, i.e., $\mathcal{O}(N!)$. In addition, each step involves an iterative convergence procedure, whose complexity is unknown. Combined with an expensive metric, the centroid is essentially intractable for datasets, which are sufficiently large to be interesting. This motivates us to investigate simpler algorithms that compute means in the Euclidean case.

2.4.2 Birkhoff shortening

Another method for computing means is given by Birkhoff curve shortening, which is used in metric and differential geometry to generate closed geodesic curves. Given a closed curve $\gamma\colon S^1 \to X$ from the unit circle into a (locally) geodesic metric space X, sample the curve by picking N points $z_1, \ldots, z_N$ on S^1 and consider their images $x_i = \gamma(z_i)$, setting $A = \{x_i | i = 1, \ldots, N\}$. Start an iterative process by replacing each point x_i by the midpoint x_i^1 of the geodesic connecting x_i to x_{i+1}, where we define $x_{N+1} = x_1$. Together with the geodesic segments we now have a new closed curve; see fig. 2.6 for an illustration. This process can be continued until convergence, which is ensured by Theorem 2.5 below. Note that when A contains three elements, the centroid procedure coincides with the Birkhoff shortening, see also fig. 2.5.

In Euclidean space it is easy to see that this process converges to the mean of the point set A. Trinh and Kimia [TK10, Conjecture 1] conjecture that the same holds for a wide range of spaces; we are less optimistic and conjecture that the claim holds locally at generic points for tree-space with the QED metric.

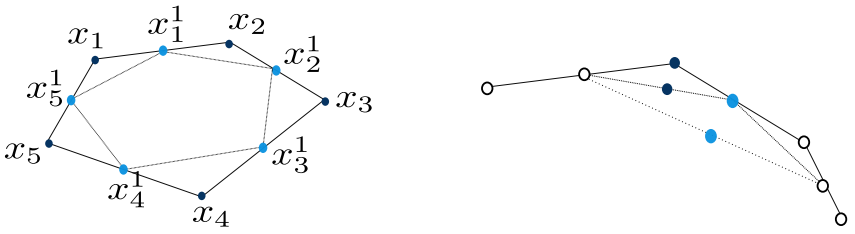


Figure 2.6: Left: Birkhoff curve shortening: Replace a sampled curve by new samples generated at the midpoint of each segment. The process converges to a closed geodesic. Right: By collecting midpoints from every second iteration (in colors), we get two Cauchy sequences, which must converge since the space is complete. It follows that the procedure converges to a closed curve.

Theorem 2.5 *Let (X, d) be a complete, geodesic metric space and suppose given an ordered set of points $A = \{x_1, \ldots, x_N\}$ in (X, d). The Birkhoff shortening process converges to a closed geodesic, and if the space (X, d) does not have closed geodesics, then the Birkhoff shortening process converges to a single point.*

A simpler form of this theorem is stated without proof by Trinh and Kimia [TK10, Proposition 1]; we give the proof in the case of a geodesic metric space, as the proof also sheds some light on the difficulties and potential dangers.

Proof of Theorem 2.5: We first note that each step in the Birkhoff shortening process *will* make the loop in question shorter, assuming that the current curve is not a closed geodesic. To see this, note that

$$\begin{aligned} \sum d(x_i, x_{i+1}) &= \sum \big(d(x_i, x_i^1) + d(x_i^1, x_{i+1})\big) \\ &= \sum \big(d(x_i^1, x_{i+1}) + d(x_{i+1}, x_{i+1}^1)\big) \\ &\geq \sum d(x_i^1, x_{i+1}^1), \end{aligned} \tag{2.6}$$

where the first equality comes from the fact that the points x_i^1 are midpoints of geodesic segments, the second equality is a rearrangement of terms, and the last inequality comes from the triangle inequality; see fig. 2.6. This shows that Birkhoff shortening will *not* make the curve longer. If the original loop is not a geodesic curve, then at one of the points x_i, the loop is not a local geodesic. The curve connecting the midpoints before and after x_i is not a geodesic, and hence, replacing with a geodesic *will* create a strictly shorter loop.

Next, we need to show that the process converges. The lengths of the consecutive curves form a decreasing sequence of non-negative real numbers, which must converge towards some length l. Moreover, the odd/even midpoint sequences shown in fig. 2.6 are Cauchy (follows from the $CAT(0)$ criterion) and must converge, so the sequence of loops converges to a new loop.

Finally, since we have already shown that for non-geodesic loops, Birkhoff shortening *will* make the curve shorter, the limit loop must be a closed geodesic. But then, if the space does not have closed geodesics, Birkhoff shortening will converge towards *one* point. □

Corollary 2.7 *Since they have unique geodesics [BH99, Proposition II.1.4], $CAT(0)$ spaces cannot have closed geodesics. Hence, the Birkhoff shortening*

procedure converges to a point in $CAT(0)$ spaces. **In particular, it converges for sets of tree-like shapes with sufficiently small diameter.**

We have shown that for any given initial order on a sufficiently bounded set of tree-like shapes, the Birkhoff shortening procedure converges to a point, but we do not know whether any two orders give the same point. For general $CAT(0)$ spaces, we doubt that this is true. Consider the Birkhoff shortening process shown in fig. 2.6 in a space whose curvature varies strongly (although staying non-positive). The shortening process would be highly asymmetric and it is unlikely that different set orders would give the same mean. This also gives potential problems with one of the algorithms used for TED-means by Trinh and Kimia, where the set order is permuted in every iteration in order to speed up convergence. However, in tree-space, the local structure is nice almost everywhere, being either flat or an intersection of flat regions. Hence, we believe that the Birkhoff shortening mean may well be independent of its initial order for sets of tree-like shapes in the QED metric.

The computational complexity of each step in the Birkhoff shortening algorithm is $\mathcal{O}(N)$ times the complexity of finding geodesic midpoints. This makes the Birkhoff shortening procedure suitable for computing means when distances are expensive to compute. The iterative nature of the procedure does, however, make it vulnerable to accumulation of numerical noise and approximation errors.

2.4.3 Weighted midpoints

One of the simplest algorithms for computing the mean of a set in $\mathbb{R}^n$ is based on the following simple observation:

If we denote by $m(A)$ the mean of the finite subset $A = \{x_1, \ldots, x_N\}$, then

$$m(A) = \frac{x_i + (N-1)m(A \setminus \{x_i\})}{N}. \tag{2.8}$$

To see that eq. 2.8 holds for $A \subset \mathbb{R}^n$, just rewrite the equation analytically in the case $i = 1$:

$$\frac{x_1 + \ldots + x_N}{N} = \frac{x_1 + (N-1) \cdot \frac{x_2 + \ldots + x_N}{N-1}}{N}. \tag{2.9}$$

This indicates a recursive procedure for finding the mean of A in more general spaces: The mean w_1 of $\{x_1, x_2\}$ is the midpoint of the geodesic connecting x_1 to x_2. The mean w_2 of $\{x_1, x_2, x_3\}$ is the point on the geodesic from $w_1 = m(x_1, x_2)$ to x_3 that sits $1/3$ along the way, etc. This is a finite procedure, whose result we call the *weighted midpoints* mean. See also fig. 2.7.

As the Birkhoff shortening mean, the weighted midpoints procedure also depends on an initial set order, and it is not clear whether different orders give the same results. Again, we believe that the nice local structure of tree-space may be enough to secure independence of initial order.

2.5 Experiments

We now experimentally compare the different approaches on datasets of tree-structured shapes in the space of tree-like shapes [FLL$^+$10] endowed with the

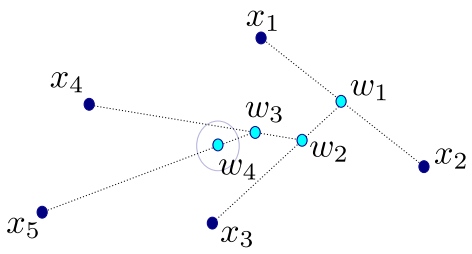


Figure 2.7: The weighted midpoints construction.

QED metric. Specifically, we test the algorithms on three different datasets of varying difficulty and size.

The QED geodesics and midpoints were computed using Algorithm 1 from the article by Feragen et al. [FLL$^+$10] on depth 3 trees, leaving space for 1 or 2 structural changes. This is an approximative algorithm, where the geodesics might pass through trees of depth higher than 3 in order to make the structural changes. Whenever the midpoints are of depth > 3, they are cut off at depth 3 in order to initiate the next iteration. This introduces some numerical errors, which may accumulate in long iterative procedures.

2.5.1 Synthetic data

The first test set consists of synthetic planar trees which are designed to test the system's ability to cope with pairs of bifurcations that are close to forming trifurcations. The whole dataset is shown in fig. 2.8. Additional figures and movies illustrating the iterative processes are found in the supplementary material.

Small dataset. First, all three different algorithms were ran on a smaller set consisting of the 4 synthetic planar trees shown in the top row of fig. 2.8. Since the weighted midpoints and Birkhoff shortening algorithms potentially depend on the order of the dataset, they were ran several times with different randomly selected initial orders. In fig. 2.9a we see the results of the weighted midpoints algorithm for all possible orders on the dataset plotted together. In fig. 2.9b, we see the points in the Birkhoff shortening 5th iteration on the same dataset. In fig. 2.9c we see the result of the centroid algorithm. We clearly see that the three algorithms give qualitatively very similar results, also with the different initial orders used in the weighted midpoints and Birkhoff shortening algorithms.

In order to experimentally check whether the found centroid c actually minimizes the function Φ from eq. 2.1, we selected 100 tree-shapes by adding random normal distributed noise to the centroid edges, and evaluated Φ at each point. The smallest value of Φ was found at the centroid, indicating that the found centroid actually is a mean.

Full dataset. The weighted means and Birkhoff shortening algorithms were also tested on the whole dataset in fig. 2.8, again using different orders. The results are found in fig. 2.10 and fig. 2.11. The centroid was left out of this experiment, since already here, the complexity is too demanding.

Although the Birkhoff shortening and weighted midpoints algorithms depend

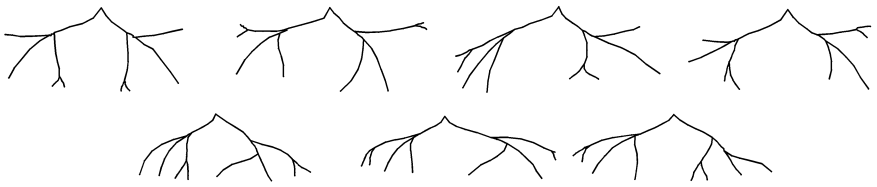


Figure 2.8: Seven synthetic planar trees.

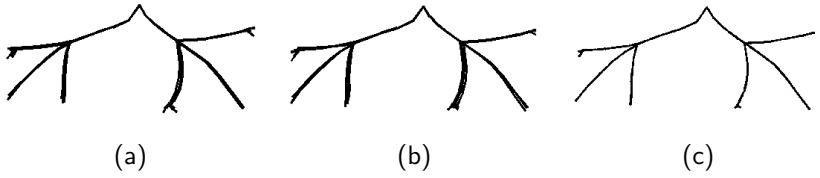


Figure 2.9: Comparison of algorithms on a set of four synthetic planar trees. (a) Results of weighted midpoints on the first four trees from fig. 2.8. The algorithm was run for all different initial orders and the results are plotted on top of each other. (b) Result of Birkhoff shortening after 5 iterations, starting from the shown set order on the first four trees from fig. 2.8. The 5th iteration points are plotted on top of each other. (c) Result of the centroid algorithm on the first four trees from fig. 2.8.

on the order of the dataset, we clearly see that the attained results are robust with respect to varying initial order. More importantly, we see that the results of the faster methods are practically identical to the centroid, which has good theoretical properties, but is expensive to compute. This is very comforting as it indicates that we can compute usable means with the Birkhoff shortening and weighted midpoints algorithms. In the remainder of this paper, we will only use these algorithms as the centroid is computationally too demanding.

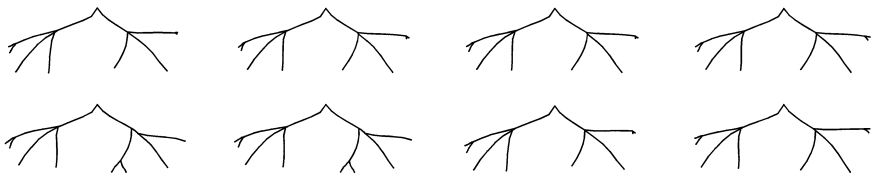


Figure 2.10: Weighted midpoints results for nine eight different initial orders on the dataset shown in fig. 2.8.

Figure 2.11: Birkhoff shortening results for three different orders on the dataset shown in fig. 2.8.

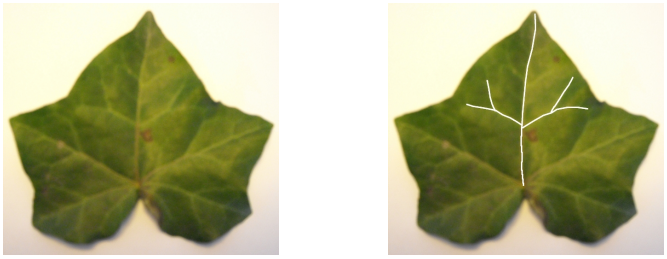


Figure 2.12: An ivy leaf, with a subtree of its vascular tree.

2.5.2 Leaf morphology

Our first example of natural tree-like structures comes from biology. In botany, tree-like structures are found as vasculatures in leaves, and are studied in order to understand leaf morphology [FA60]. These structures form excellent proof-of-concept examples, as the tree-structures are necessarily planar, and hence the branches are ordered nicely from left to right.

We extract vascular structures for a set of 10 ivy leaves, see fig. 2.12, giving the planar trees shown in fig. 2.13. Using weighted midpoints for six randomly chosen dataset orders we obtain the mean trees showed in fig. 2.14. Again, the weighted midpoints mean trees initiated with different initial orders look nearly identical. Similarly, in fig. 2.15, we see 12th iteration Birkhoff shortening mean trees for a random initial order, which are also very similar to the weighted midpoints results. This is a clear indication that the two algorithms compute the same mean.

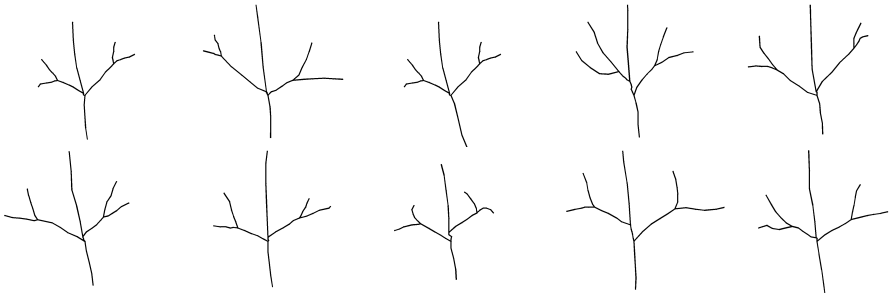


Figure 2.13: Vascular structures from 10 ivy leaves.

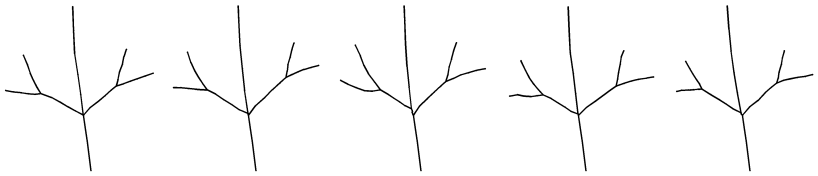


Figure 2.14: weighted midpoints results for the leaves shown in fig. 2.13 for five random dataset orders.

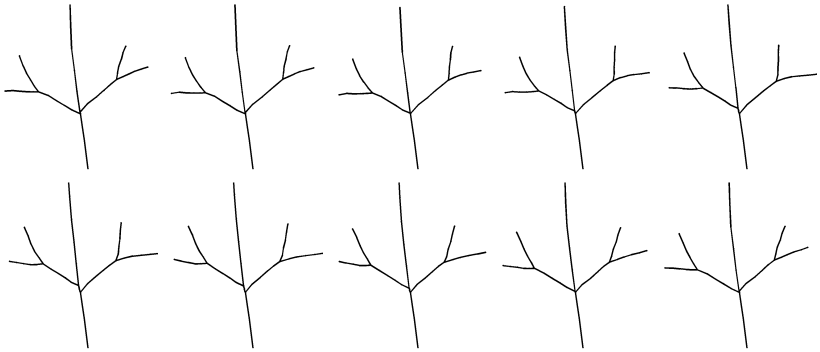


Figure 2.15: The Birkhoff shortening 12th iteration results for the leaves shown in fig. 2.13 for a random dataset order.

2.5.3 Airway tree-shape modeling

The study of tree-like shapes is strongly motivated by their presence in human anatomy, where they appear as delivery systems for fluids and air. Birkhoff shortening and weighted midpoints means are computed for a dataset consisting of 10 $3D$ airway trees extracted from CT scans [LSA$^+$10]. The trees represent the centerlines of the first four generations of the human airway tree as shown in fig. 2.16. The trees are aligned at the endpoint of the root branch (trachea). As with the planar trees, the two algorithms provide very similar means. Both algorithms, and Birkhoff shortening in particular, seem robust even for $3D$ trees.

2.6 Discussion and conclusion

We have studied the concept of means in non-Euclidean spaces; generally in $CAT(0)$ spaces and specifically in spaces of tree-like shapes. We have shown that means exist in $CAT(0)$ spaces; a result which tells us that means exists locally in spaces of tree-like shapes. The generality of the result allows it to be transferred to other settings. In particular, our results should transfer to the space of attributed graphs defined by Jain and Obermayer [JO10].

Usually, means can be found in non-Euclidean spaces using standard optimization techniques such as gradient descent. The computational complexity of the QED metric, however, makes this approach infeasible for trees. We consider three different algorithms for computing means in Euclidean space: the cen-

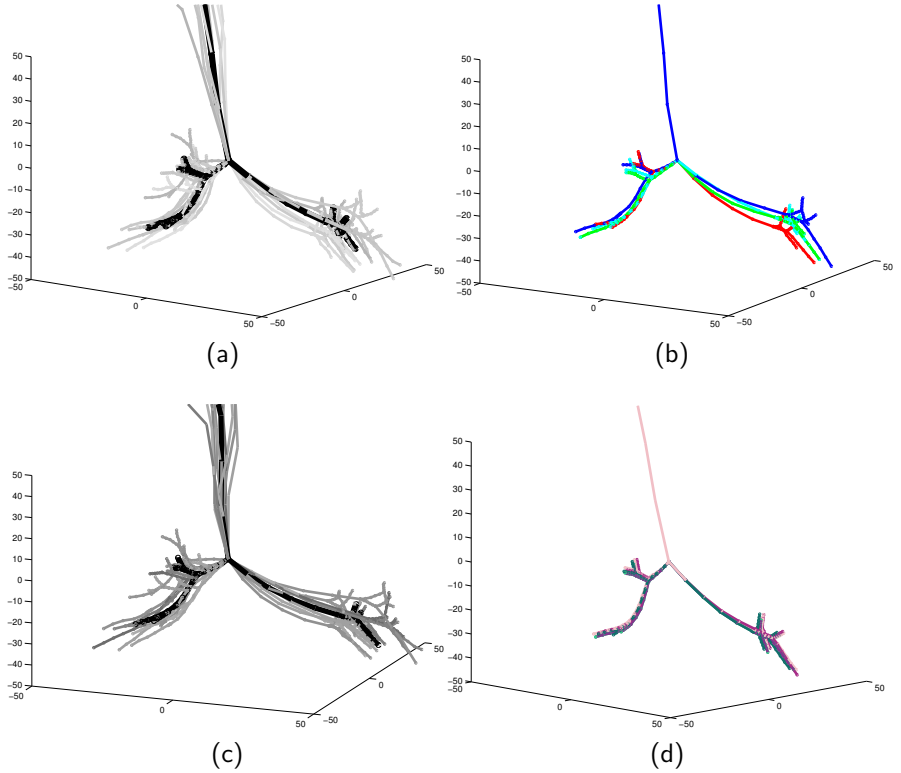


Figure 2.16: (a) The ten airway trees in the dataset are shown in gray; in black is one of the weighted midpoints means. (b) Five weighted midpoints means computed from different initial orders. (c) The ten airway trees in the dataset are shown in gray; in black is the Birkhoff shortening mean. (d) Five Birkhoff shortening means computed from different initial orders.

troid, Birkhoff shortening and weighted midpoints. As these rely on on dividing geodesics they are readily generalized to non-Euclidean spaces.

From a theoretical point of view, the centroid is the best definition of mean shape. It exists, is unique and is, by definition, invariant of the order of the dataset. Moreover, numerical experiments indicate that it coincides with the mean. The algorithm, however, has complexity $\mathcal{O}(N!)$, which makes it too expensive to be of practical use. On the other hand, the Birkhoff shortening and weighted midpoints means are not quite as nice; while they converge to a single point, this point may depend on the order of the data. This is not ideal. However, experimental results indicate that, up to numerical and approximative noise, different orders actually give the same means for both methods. Even better: these means seem to coincide with the well-defined centroid. On this basis, we conjecture that for tree-structured data, all algorithms compute the same mean tree. As to which method works the best, the simplest seems to win. Birkhoff shortening is an iterative procedure, which makes it slow and vulnerable to accumulating errors. The weighted midpoints mean, on the other hand, comes out as a robust and efficient way of computing mean trees.

We have, thus, provided a practical algorithm for computing mathematically well-defined "average trees"; something that has not been presented elsewhere in the literature. This can potentially serve as a much-needed tool in medical image analysis, where airways and vascularization systems serve as reference structures in the human body. Shape statistics on these structures may provide new insight into how diseases such as COPD (smokers lung) affects the geometry of human anatomy.

Acknowledgements

This work is partly funded by the Lundbeck foundation, the Danish Council for Strategic Research (NABIIT projects 09 − 065145 and 09 − 061346) and the Netherlands Organisation for Scientific Research (NWO). We thank Dr. J.H. Pedersen from the Danish Lung Cancer Screening Trial (DLCST) for providing the CT scans.

We thank Marco Loog and Melanie Ganz for helpful discussions during the preparation of the article.

Bibliography

[APW+09] B. Aydin, G. Pataki, H. Wang, E. Bullitt, and J. S. Marron. A principal component analysis for trees. *Ann. Appl. Statistics*, (4):1597–1615, 2009.

[BH99] Martin R. Bridson and André Haefliger. *Metric spaces of non-positive curvature*. Springer-Verlag, 1999.

[BHV01] Louis J. Billera, Susan P. Holmes, and Karen Vogtmann. Geometry of the space of phylogenetic trees. *Adv. in Appl. Math.*, 27(4):733–767, 2001.

[BLWH06] T. Bülow, C. Lorenz, R. Wiemker, and J. Honko. Point based methods for automatic bronchial tree matching and labeling. In *SPIE Med. Im.*, volume 6143, pages 225–234, 2006.

[BWL+09] Xiang Bai, Xinggang Wang, Longin Jan Latecki, Wenyu Liu, and Zhuowen Tu. Active skeleton for non-rigid object detection. In *ICCV*, pages 575–582, 2009.

[CAM+05] A. Charnoz, V. Agnus, G. Malandain, L. Soler, and M. Tajine. Tree matching applied to vascular system. *GbRPR, LNCS*, 3434:183–192, 2005.

[CFM01] C. Chalopin, G. Finet, and I. E. Magnin. Modeling the 3D coronary tree for labeling purposes. *MIA*, 5:301–315, 2001.

[DPS+09] M. Demirci, Bram Platel, Ali Shokoufandeh, Luc Florack, and Sven Dickinson. The representation and matching of images using top points. *Journal of Mathematical Imaging and Vision*, 35:103–116, 2009.

[DSD09] F. Demirci, A. Shokoufandeh, and S.J. Dickinson. Skeletal shape abstraction from examples. *TPAMI*, 31(5):944–952, 2009.

[Dug66] James Dugundji. *Topology*. Allyn and Bacon Inc., 1966.

[FA60] Adriance S. Foster and Howard J. Arnott. Morphology and dichotomous vasculature of the leaf of kingdonia uniflora. *American Journal of Botany*, 47(8):pp. 684–698, 1960.

[FLL+10] A. Feragen, F. Lauze, P. Lo, M. de Bruijne, and M. Nielsen. Geometries in spaces of treelike shapes. In *ACCV*, pages 671–684, 2010.

[FLN10] A. Feragen, F. Lauze, and M. Nielsen. Fundamental geodesic deformations in spaces of treelike shapes. In *Proc ICPR*, 2010.

[FLPJ04] P. Thomas Fletcher, Conglin Lu, Stephen M. Pizer, and Sarang Joshi. Principal geodesic analysis for the study of nonlinear statistics of shape. *TMI*, 23:995–1005, 2004.

[FVS+10] M. Ferrer, E. Valveny, F. Serratosa, K. Riesen, and H. Bunke. Generalized median graph computation by means of graph embedding in vector spaces. *Pattern Recognition*, 43(4):1642 – 1655, 2010.

[GK03] Peter J. Giblin and Benjamin B. Kimia. On the local form and transitions of symmetry sets, medial axes, and shocks. *IJCV*, 54(1-3):143–156, 2003.

[Gro87] M. Gromov. Hyperbolic groups. In *Essays in group theory*, volume 8 of *Math. Sci. Res. Inst. Publ.*, pages 75–263. 1987.

[HHM10] S. Huckemann, T. Hotz, and A. Munk. Intrinsic shape analysis: geodesic PCA for Riemannian manifolds modulo isometric Lie group actions. *Statist. Sinica*, 20(1):1–58, 2010.

[JO09] Brijnesh J. Jain and Klaus Obermayer. Structure spaces. *J. Mach. Learn. Res.*, 10:2667–2714, 2009.

[JO10] B. Jain and K. Obermayer. Large sample statistics in the domain of graphs. In *SSPR*, volume 6218 of *LNCS*, pages 690–697. 2010.

[Kar77] H. Karcher. Riemannian center of mass and mollifier smoothing. *Communications on Pure and Applied Mathematics*, 30(5):509–541, 1977.

[Ken84] David G. Kendall. Shape manifolds, Procrustean metrics, and complex projective spaces. *Bull. London Math. Soc.*, 16(2):81–121, 1984.

[KF05] Arjan Kuijper and Luc M. J. Florack. Using catastrophe theory to derive trees from images. *J. Math. Imaging Vision*, 23(3):219–238, 2005.

[KKNN06] J.N. Kaftan, A.P. Kiraly, D.P. Naidich, and C.L. Novak. A novel multi-purpose tree and path matching algorithm with application to airway trees. In *SPIE*, volume 6143 I, 2006.

[KSDD03] Yakov Keselman, Ali Shokoufandeh, M. Fatih Demirci, and Sven Dickinson. Many-to-many graph matching via metric embedding. In *In CVPR*, volume 1, pages 850–857, 2003.

[KSK01] Philip N. Klein, Thomas B. Sebastian, and Benjamin B. Kimia. Shape matching using edit-distance: an implementation. In *Proceedings of the twelfth annual ACM-SIAM symposium on Discrete algorithms*, SODA '01, pages 781–790, 2001.

[KTSK00] Philip Klein, Srikanta Tirthapura, Daniel Sharvit, and Ben Kimia. A tree-edit-distance algorithm for comparing simple, closed shapes. In *In SODA*, pages 696–704, 2000.

[LSA+10] Pechin Lo, Jon Sporring, Haseem Ashraf, Jesper J.H. Pedersen, and Marleen de Bruijne. Vessel-guided airway tree segmentation: A voxel classification approach. *Medical Image Analysis*, 14(4):527–538, 2010.

[MBK+09] V. Megalooikonomou, M. Barnathan, D. Kontos, Pe. R. Bakic, and A. D.A. Maidment. A representation and classification scheme for tree-like structures in medical images: Analyzing the branching pattern of ductal trees in x-ray galactograms. *TMI*, 28(4):487–493, 2009.

[MKS+09] J. H. Metzen, T. Kröger, A. Schenk, S. Zidowitz, H-O. Peitgen, and X. Jiang. Matching of anatomical tree structures for registration of medical images. *Im. Vis. Comp.*, 27:923–933, 2009.

[MM04] Peter W. Michor and David Mumford. Riemannian geometries on spaces of plane curves. *J. Eur. Math. Soc. (JEMS*, 8:1–48, 2004.

[OP11] Megan Owen and J. Scott Provan. A fast algorithm for computing geodesic distances in tree space. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 8:2–13, 2011.

[PAD+09] Jesper Pedersen, Haseem Ashraf, Asger Dirksen, Karen Bach, Hanne Hansen, Phillip Toennesen, Hanne Thorsen, John Brodersen, Birgit Skov, Martin Døssing, Jann Mortensen, Klaus Richter, Paul Clementsen, and Niels Seersholm. The Danish randomized lung cancer CT screening trial - overall design and results of the prevalence round. *J Thorac Oncol*, 4(5):608–614, May 2009.

[RB09a] Kaspar Riesen and Horst Bunke. Approximate graph edit distance computation by means of bipartite graph matching. *Image and Vision Computing*, 27(7):950 – 959, 2009. 7th IAPR-TC15 Workshop on Graph-based Representations (GbR 2007).

[RB09b] Kaspar Riesen and Horst Bunke. Graph classification based on vector space embedding. *International Journal of Pattern Recognition & Artificial Intelligence*, 23(6):1053 – 1081, 2009.

[SBK+10] D. Smeets, P. Bruyninckx, J. Keustermans, D. Vandermeulen, and P. Suetens. Robust matching of 3d lung vessel trees. In *MICCAI 2010, Proc 3rd Intl WS Pulm. Im. Anal.*, pages 61–70, 2010.

[SKK02] Thomas Sebastian, Philip Klein, and Benjamin Kimia. Shock-based indexing into large shape databases. In *Computer Vision â€" ECCV 2002*, volume 2352 of *Lecture Notes in Computer Science*, pages 83–98. 2002.

[SKK04] T. B. Sebastian, P.N. Klein, and B.B. Kimia. Recognition of shapes by editing their shock graphs. *TPAMI*, 26(5):550–571, 2004.

[SLC+02] Thorsten Schlathölter, Cristian Lorenz, Ingwer C. Carlsen, Steffen Renisch, and Thomas Deschamps. Simultaneous segmentation and tree reconstruction of the airways for virtual bronchoscopy. volume 4684, pages 103–113. SPIE, 2002.

[SLD+11] L. Sørensen, P. Lo, A. Dirksen, J. Petersen, and Marleen de Bruijne. Dissimilarity-based classification of anatomical tree structures. In GÃ¡bor SzÃ©kely and Horst Hahn, editors, *Information Processing in Medical Imaging*, Lecture Notes in Computer Science. Springer, 2011.

[TH03] Andrea Torsello and Edwin R. Hancock. Computing approximate tree edit distance using relaxation labeling. *Pattern Recognition Letters*, 24(8):1089 – 1097, 2003. Graph-based Representations in Pattern Recognition.

[TK10] N.H. Trinh and B.B. Kimia. Learning prototypical shapes for object categories. In *CVPR Workshops*, pages 1–8, 2010.

[TMP+05] J. Tschirren, G. McLennan, K. Palágyi, E. A. Hoffman, and M. Sonka. Matching and anatomical labeling of human airway tree. *TMI*, 24(12):1540–1547, 2005.

[vGBvR08] Bram van Ginneken, Wouter Baggerman, and Eva van Rikxoort. Robust segmentation and anatomical labeling of the airway tree from thoracic CT scans. In *Medical Image Computing and Computer-Assisted Intervention â€“ MICCAI 2008*, pages 219–226, 2008.

[VMS05] Rene Vidal, Yi Ma, and Shankar Sastry. Generalized principal component analysis (gpca). *IEEE Trans. Pattern Anal. Mach. Intell.*, 27:1945–1959, December 2005.

[WDE+09] G. R. Washko, T. Dransfield, R. S. J. Estepar, A. Diaz, S. Matsuoka, T. Yamashiro, H. Hatabu, E. K. Silverman, W. C. Bailey, and J. J. Reilly. Airway wall attenuation: a biomarker of airway disease in subjects with COPD. *J Appl Physiol*, 107(1):185–191, 2009.

[Wei09] E. R. Weibel. What makes a good lung? *Swiss Med. Weekly*, (139(27-28)):375–386, 2009.

[WM07] Haonan Wang and J. S. Marron. Object oriented data analysis: sets of trees. *Ann. Statist.*, 35(5):1849–1873, 2007.

[ZSS92] Kaizhong Zhang, Rick Statman, and Dennis Shasha. On the editing distance between unordered labeled trees. *Inf. Process. Lett.*, 42(3):133–139, 1992.