Dept. of Computer Science, University of Copenhagen

Towards a theory of statistical tree-shape analysis

Banff workshop on geometry for anatomy

Aasa Feragen, Francois Lauze, Marleen de Bruijne, **Mads Nielsen**

Dept. of Computer Science, University of Copenhagen

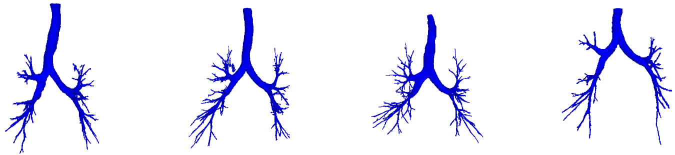
Towards a theory of statistical tree-shape analysis

Banff workshop on geometry for anatomy

Aasa Feragen, Francois Lauze, Marleen de Bruijne, Mads Nielsen

Motivation

What does the average human airway tree look like?

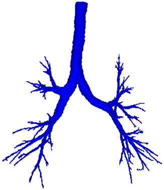


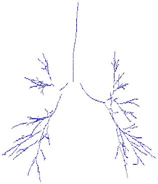
Nobody knows! There are no tools available for doing statistics on airway trees!

Motivation

The airway tree can be described as a combination of

- tree topology (connectivity / combinatorics)
- geometry (branch shape)

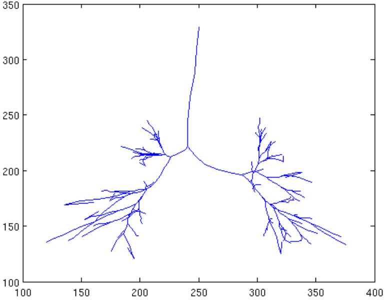




Motivation

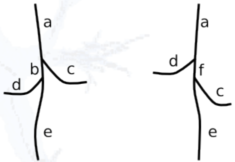
Tree:

- vertices
- edges connecting vertices
- root
- order



Motivation

So why don't you just collect the edge information in a long vector and compute averages? Consider the *rather similar* trees:



which are represented by the *rather different* vectors

$$(a, b, c, d, e) \text{ and } (a, d, f, e, c).$$

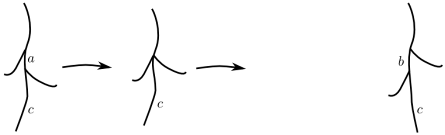
We need methods which can gracefully handle topological differences.

Tree distance example: Tree edit distance (TED)

- TED is a classical, algorithmic distance
- tree-space with TED is a ”funny space”
- dist(T_1, T_2) is the minimal total cost of changing T_1 into T_2 through three basic operations:
- Remove edge, add edge, deform edge.

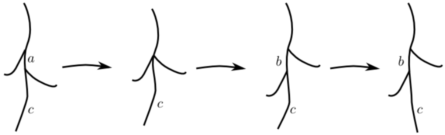
Tree distance example: Tree edit distance (TED)

- TED is a classical, algorithmic distance
- tree-space with TED is a ”funny space”
- dist(T_1, T_2) is the minimal total cost of changing T_1 into T_2 through three basic operations:
- Remove edge, add edge, deform edge.



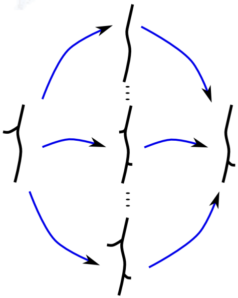
Tree distance example: Tree edit distance (TED)

- TED is a classical, algorithmic distance
- tree-space with TED is a "funny space"
- dist(T_1, T_2) is the minimal total cost of changing T_1 into T_2 through three basic operations:
- Remove edge, add edge, deform edge.



Tree distance example: Tree edit distance (TED)

- Almost all geodesics between pairs of trees are non-unique (infinitely many).



- Then what is the average of two trees? Many!
- TED is *not* suitable for statistics.

madsn@diku.dk

Tree distance example: Tree edit distance (TED)

Most state-of-the-art approaches to distance measures and statistics on tree- and graph-structured data *are* based on TED!

- Wang and Marron: Object oriented data analysis: sets of trees. Annals of Statistics 35 (5), 2007.
- Ferrer, Valveny, Serratosa, Riesen, Bunke: Generalized median graph computation by means of graph embedding in vector spaces. Pattern Recognition 43 (4), 2010.
- Riesen and Bunke: Approximate Graph Edit Distance by means of Bipartite Graph Matching. Image and Vision Computing 27 (7), 2009.
- Trinh and Kimia, Learning Prototypical Shapes for Object Categories. CVPR workshops 2010.

Tree distance example: Tree edit distance (TED)

The problems can be "solved" by choosing specific geodesics.
OBS! Geometric methods can no longer be used for proofs, and one risks choosing problematic paths.

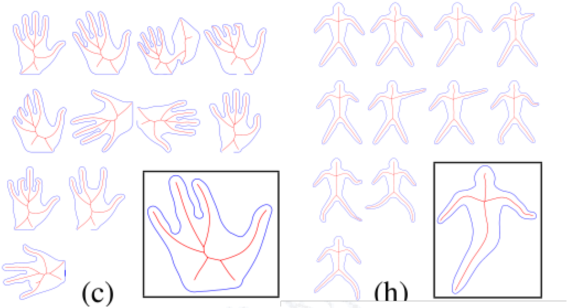


Figure: Trinh and Kimia (CVPR workshops 2010) compute average shock graphs using TED with the simplest possible choice of geodesics.

Modeling trees – the ideal model

The model we are looking for

- a geodesic metric structure, with good uniqueness properties

The model we are looking for

- a geodesic metric structure, with good uniqueness properties
- geodesics corresponding to natural deformations

The model we are looking for

- a geodesic metric structure, with good uniqueness properties
- geodesics corresponding to natural deformations
- averages and modes of variation, with good uniqueness properties

madsn@diku.dk

The model we are looking for

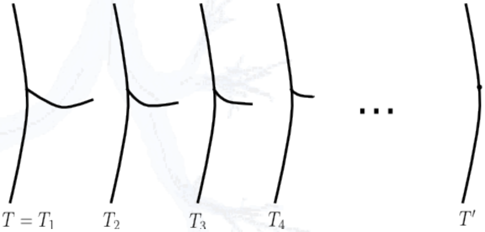


Figure: Tolerance of structural noise.

Building a space of tree-like shapes [1]

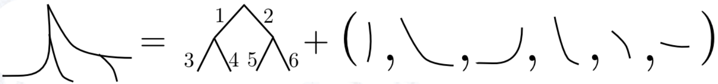
[1] A. Feragen, F. Lauze, P. Lo, M. de Bruijne, M. Nielsen, ACCV2010

Tree representation

How to represent tree-like shapes mathematically?

Tree-like (pre-)shape = pair $(\mathscr{T}, x)$

- $\mathscr{T} = (V, E, r, <)$ rooted, ordered/planar binary tree, describing the tree topology (combinatorics)

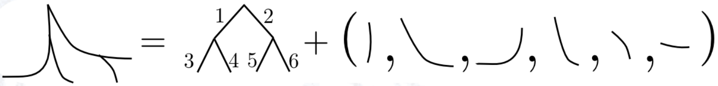


Tree representation

How to represent tree-like shapes mathematically?

Tree-like (pre-)shape = pair $(\mathscr{T}, x)$

- $\mathscr{T} = (V, E, r, <)$ rooted, ordered/planar binary tree, describing the tree topology (combinatorics)
- $x \in \prod_{e \in E} A$ a product of points in attribute space A describing edge shape



Tree representation

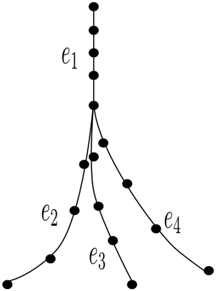
We are allowing collapsed edges, which means that

- we can represent higher order vertices
- we can represent trees of different sizes using the same combinatorial tree $\mathscr{T}$

(dotted line = collapsed edge = zero/constant attribute)

Tree representation

Edge representation through landmark points:
Edge shape space is $(\mathbb{R}^d)^n$, $d = 2, 3$.



The space of tree-like preshapes

Fix a maximal combinatorial $\mathscr{T}$. We use a finite tree; could reformulate for infinite trees.

Definition

Define the space of tree-like *pre*-shapes as the product space

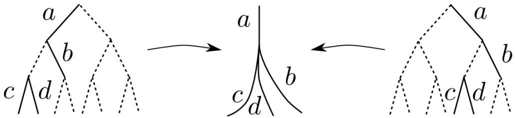
$$\prod_{e \in E} (\mathbb{R}^d)^n$$

where $(\mathbb{R}^d)^n$ is the edge shape space.

This is just a space of *pre-shapes*.

From pre-shapes to shapes

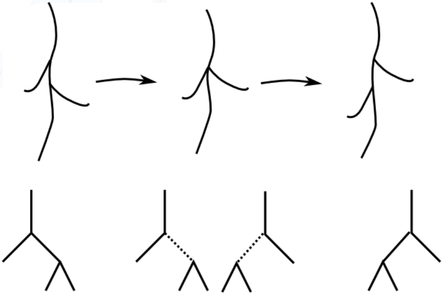
Many shapes have more than one representation



madsn@diku.dk

From pre-shapes to shapes

Not all shape deformations can be recovered as natural paths in the pre-shape space:

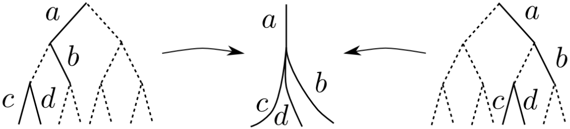


Shape space definition

- Start with the pre-shape space $X = \prod_{e \in E} (\mathbb{R}^d)^n$.

Shape space definition

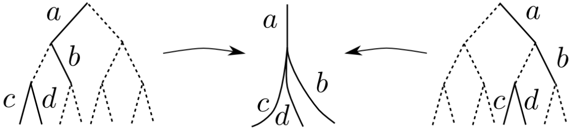
- Start with the pre-shape space $X = \prod_{e \in E} (\mathbb{R}^d)^n$.
- Define an equivalence $\sim$ by identifying points in X that represent the same tree-shape.



madsn@diku.dk

Shape space definition

- Start with the pre-shape space $X = \prod_{e \in E} (\mathbb{R}^d)^n$.
- Define an equivalence $\sim$ by identifying points in X that represent the same tree-shape.

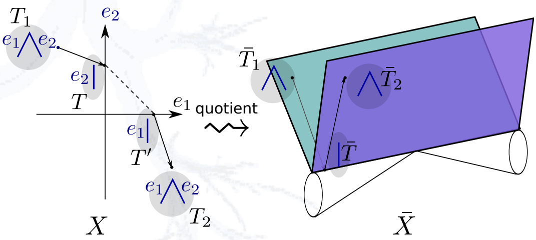


- This corresponds to identifying, or gluing together, subspaces $\{x \in X | x_e = 0 \text{ if } e \notin E_1\}$ and $\{x \in X | x_e = 0 \text{ if } e \notin E_2\}$ in X.

madsn@diku.dk

Shape space definition

- Define the space of ordered (planar) tree-like shapes $\bar{X} = X/\sim$.



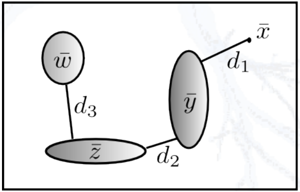
- For the landmark point shape space this is just a folded Euclidean space.

Geometries on the space of trees

Shape space metric definition

Given a metric d on the vector space $X = \prod_{e \in E} (\mathbb{R}^d)^n$ we define the quotient pseudometric $\bar{d}$ on the quotient space $\bar{X} = X/\sim$ by setting

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



$$\bar{d}(\bar{x}, \bar{w}) = d_1 + d_2 + d_3$$

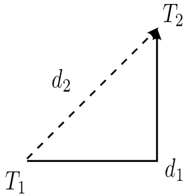
Theorem

The quotient pseudometric $\bar{d}$ is a metric on $\bar{X}$. □

Shape space metric definition

Define two metrics d_1 and d_2 on X, induced by two different product norms on the product of edge shape spaces $X = \prod_{e \in E} (\mathbb{R}^d)^n$:

$$
\begin{aligned}
l1 \text{ norm:} \quad & d_1(x, y) = \sum_{e \in E} \|x_e - y_e\| \\
l2 \text{ norm:} \quad & d_2(x, y) = \sqrt{\sum_{e \in E} \|x_e - y_e\|^2}
\end{aligned}
$$



There are metrics $\bar{d}_1$ and $\bar{d}_2$ on $\bar{X}$ induced by d_1 and d_2.

Shape space metric definition

It turns out that $\bar{d}_1$ is an old friend; namely the well-known Tree Edit Distance metric:

Theorem

The metric $\bar{d}_1$ is the TED metric on trees that "fit" into $\bar{X}$. □

Terminology

Refer to $\bar{d}_2$ as the QED (Quotient Euclidean Distance) metric.

madsn@diku.dk

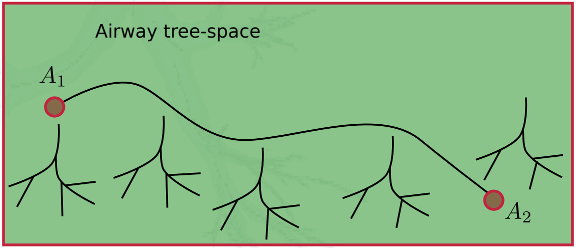
Geodesics in metric spaces

Theorem

Let $\bar{d} = \bar{d}_1$ or $\bar{d}_2$. Then $(\bar{X}, \bar{d})$ is a contractible, complete, proper geodesic space. □

How does the QED fulfill our wishes?

It defines a geodesic metric space



How does the QED fulfill our wishes?

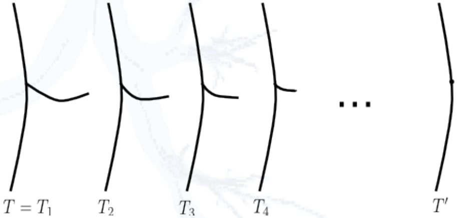
Example of a QED geodesic deformation:

Play movie

Note the tolerance of topological differences and natural deformation.

How does the QED fulfill our wishes?

Noise tolerance:

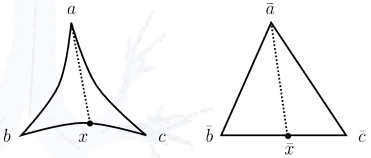


Sequences of trees with disappearing branches will converge towards trees without the same branch.

How does the QED fulfill our wishes?

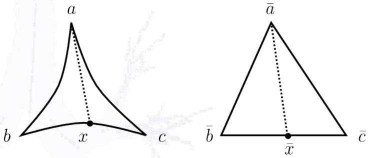
- We're doing OK so far!
- Let's return to geometry to look for uniqueness and statistical properties

Curvature in metric spaces



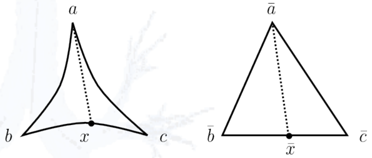
- A $CAT(0)$ space is a metric space in which geodesic triangles are "thinner" than for their comparison triangles in the plane; that is, $d(x, a) \leq d(\bar{x}, \bar{a})$.

Curvature in metric spaces



- A $CAT(0)$ space is a metric space in which geodesic triangles are "thinner" than for their comparison triangles in the plane; that is, $d(x, a) \leq d(\bar{x}, \bar{a})$.
- A space has non-positive curvature if it is locally $CAT(0)$.

Curvature in metric spaces



- A $CAT(0)$ space is a metric space in which geodesic triangles are "thinner" than for their comparison triangles in the plane; that is, $d(x, a) \leq d(\bar{x}, \bar{a})$.
- A space has non-positive curvature if it is locally $CAT(0)$.
- (Similarly define curvature bounded by κ by using comparison triangles in hyperbolic space or spheres of curvature κ.)

madsn@diku.dk

Curvature in metric spaces

Example

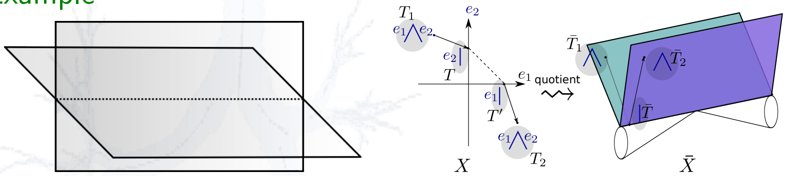


Figure: Left: $CAT(0)$ space. Right: With $A = \mathbb{R}^N$ and the QED metric; locally $CAT(0)$ except for at the origin.

Curvature in metric spaces

Example

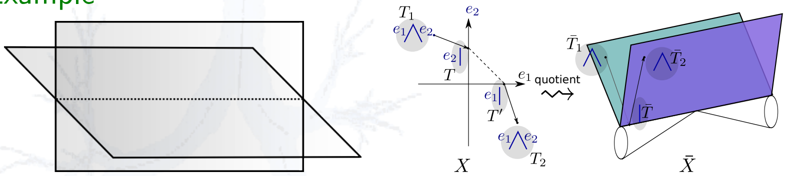


Figure: Left: $CAT(0)$ space. Right: With $A = \mathbb{R}^N$ and the QED metric; locally $CAT(0)$ except for at the origin.

Theorem (see e.g. Bridson-Haefliger)

Let (X, d) be a $CAT(0)$ space; then all pairs of points have a unique geodesic joining them. The same holds locally in $CAT(\kappa)$ spaces, $\kappa \neq 0$. □

Curvature of shape space

Theorem

- Consider $(\bar{X}, \bar{d}_2)$, shape space with the QED metric.
- At generic points, this space has non-positive curvature.
- Its geodesics are locally unique at generic points.
- At non-generic points, the curvature is unbounded. □

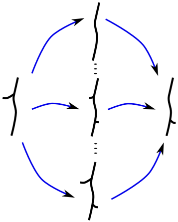
Curvature of shape space

Theorem

- Consider $(\bar{X}, \bar{d}_2)$, shape space with the QED metric.
- At generic points, this space has non-positive curvature.
- Its geodesics are locally unique at generic points.
- At non-generic points, the curvature is unbounded. □

Theorem

- Consider $(\bar{X}, \bar{d}_1)$, shape space with TED.
- It has nowhere locally unique geodesics.
- Its curvature is everywhere unbounded.



madsn@diku.dk

3D trees[2]

So far we talked about ordered (planar) tree-like shapes; what about unordered (spatial) tree-like shapes?

[2] A. Feragen, P. Lo, M. de Bruijne, M. Nielsen, F. Lauze, submitted to TPAMI

madsn@diku.dk

3D trees[2]

- Planar shapes have a given edge order.
- Unordered trees: Give a random order
- Denote by G the group of reorderings of the edges that do not alter the connectivity of the tree.
- The space of spatial/unordered trees is the space $\bar{\bar{X}} = \bar{X}/G$
- Give $\bar{\bar{X}}$ the quotient pseudometric $\bar{\bar{d}}$.
- $\bar{\bar{d}}(\bar{\bar{x}}, \bar{\bar{y}})$ corresponds to considering all possible orders on $\bar{\bar{y}}$ and choosing the order that minimizes $\bar{\bar{d}}(\bar{\bar{x}}, \bar{\bar{y}})$.

[2] A. Feragen, P. Lo, M. de Bruijne, M. Nielsen, F. Lauze, submitted to TPAMI

3D trees[2]

Theorem

- For the quotient pseudometric $\bar{\bar{d}}$ induced by either $\bar{d}_1$ or $\bar{d}_2$, the function $\bar{\bar{d}}$ is a metric and $(\bar{\bar{X}}, \bar{\bar{d}})$ is a contractible, complete, proper geodesic space.
- At non-generic points, $(\bar{\bar{X}}, \bar{\bar{d}}_2)$ has non-positive curvature.
- On the other hand, $(\bar{\bar{X}}, \bar{\bar{d}}_1)$ has everywhere unbounded curvature.
- ...so everything we proved for ordered trees, still holds. □

[2] A. Feragen, P. Lo, M. de Bruijne, M. Nielsen, F. Lauze, submitted to TPAMI

madsn@diku.dk

Statistical properties [3]

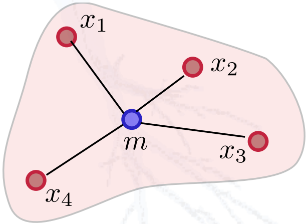
[3] A. Feragen, S. Hauberg, M. Nielsen, F. Lauze, ICCV 2011 (to appear)

Averages in the QED metric

Endow $\bar{X}$ with the QED metric $\bar{d}_2$. For a generic point $\bar{x} \in \bar{X}$, there is a radius $r_{\bar{x}}$ s.t sets $\{x_i\}_{i=1}^s$ contained in $B(\bar{x}, r_{\bar{x}})$:

Theorem

...have unique means, defined as $\operatorname{argmin} \sum d(x, x_i)^2$.

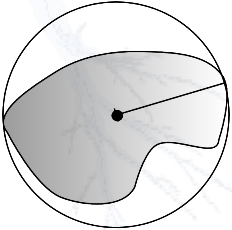


Averages in the QED metric

Endow $\bar{X}$ with the QED metric $\bar{d}_2$. For a generic point $\bar{x} \in \bar{X}$, there is a radius $r_{\bar{x}}$ s.t sets $\{x_i\}_{i=1}^{s}$ contained in $B(\bar{x}, r_{\bar{x}})$:

Theorem

...have unique circumcenters, defined as the center of the smallest sphere containing all the $\{x_i\}_{i=1}^{s}$.



madsn@diku.dk

Averages in the QED metric

Endow $\bar{X}$ with the QED metric $\bar{d}_2$. For a generic point $\bar{x} \in \bar{X}$, there is a radius $r_{\bar{x}}$ s.t sets $\{x_i\}_{i=1}^{s}$ contained in $B(\bar{x}, r_{\bar{x}})$:

Theorem (Billera, Vogtmann, Holmes)

...have unique centroids, defined by induction on $|S| = n$:

- If $|S| = 2$, then $c(S)$ is the midpoint of the geodesic between the two elements of S.
- If $|S| = n > 2$ and we have defined $c(S')$ for all S' with $|S'| < n$, then denote by $c^1(S)$ the set $\{c(S')|S' \subset S, |S'| = n-1\}$ and denote by $c^k(S) = c^1(c^{k-1}(S))$ when $k > 1$.
- If $c^k(S) \to p$ for some $p \in \bar{X}$ as $k \to \infty$, then $c(S) = p$ is the centroid of S. □

madsn@diku.dk

Averages in the QED metric

Synthetic data:

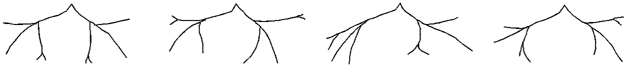


Figure: A small set of synthetic planar tree-shapes.

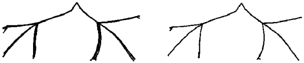


Figure: Left: Mean shape. Right: Centroid shape.

These choices of "average" give rather similar results.

Averages in the QED metric

Leaf vasculature data:

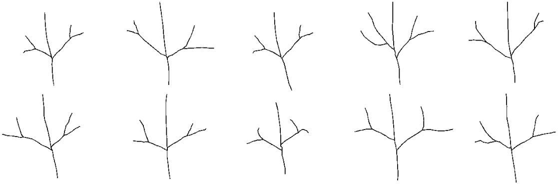


Figure: A set of vascular trees from ivy leaves form a set of planar tree-shapes.

Figure: a): The vascular trees are extracted from photos of ivy leaves. b) The mean vascular tree.

Airway shape modeling

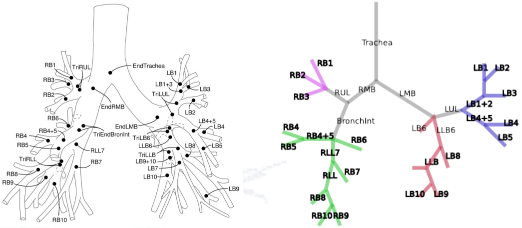


Figure: Left figure borrowed from Tschirren et al. [4]

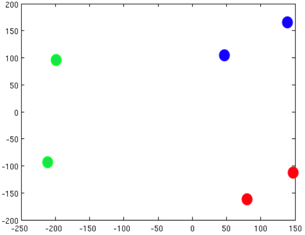
- Combinatorial structure of airway tree is somewhat fixed, for anatomical reasons.
- There are topological differences, making both global and local comparison difficult.

[4]Tschirren et al: Matching and anatomical labeling of human airway tree, TMI 2005

Airway shape modeling

Experiment 1: Compute approximate ($k \leq 3$) geodesic distances between six airways, two images from each of three different people. We can clearly distinguish patients: [4]

	P(1,1)	P(1,2)	P(2,1)	P(2,2)	P(3,1)	P(3,2)
P(1,1)	0	309.09	437.58	452.62	375.40	378.19
P(1,2)	309.09	0	435.11	402.71	400.41	349.41
P(2,1)	437.58	435.11	0	400.91	448.45	392.69
P(2,2)	452.62	402.71	400.91	0	456.69	411.24
P(3,1)	375.40	400.41	448.45	456.69	0	324.43
P(3,2)	378.19	349.41	392.69	411.24	324.43	0



[4] A. Feragen, P. Lo, M. de Bruijne, F. Lauze and M. Nielsen, ACCV2010.

Airway shape modeling

Experiment 2: Combine geodesic deformations with a voting scheme to induce anatomical branch labeling on 20 noisy airways from 15 subjects. Average correct labeling rate of 83%. [4]

CASE	21	22	23	24	25	26	27	28	29	30
% correct	75	88.2	92.9	80	77.8	86.7	88.9	94.4	66.7	89.5
CASE	31	32	33	34	35	36	37	38	39	40
% correct	90	76.5	88.9	100	83.3	78.9	66.7	80	30	76.5

Figure: Correct labeling quotas for the different airway trees.

[4] A. Feragen P. Lo, V. Gorbunova, A. Dirksen, J. Reinhardt and M. de Bruijne, submitted.

Airway shape modeling

Experiment 4: The mean upper airway tree[4]

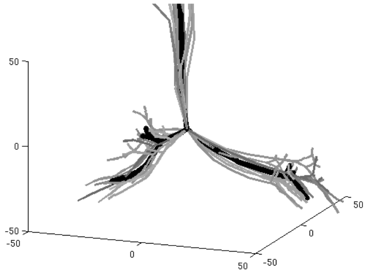


Figure: A set of upper airway tree-shapes along with their mean tree-shape.

[4] A. Feragen, S. Hauberg, M. Nielsen and F. Lauze, ICCV 2011

Airway shape modeling

Experiment 5:

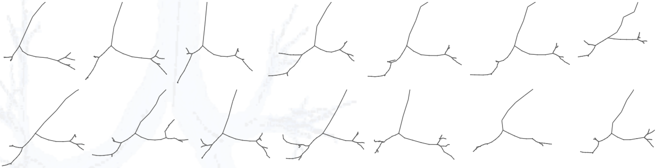


Figure: A set of upper airway tree-shapes (projected).[4]

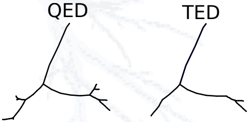


Figure: The QED and TED (algorithm by Trinh and Kimia) means.

[4] with P. Lo, M. de Bruijne, M. Nielsen, F. Lauze, submitted to TPAMI

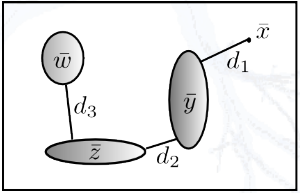
madsn@diku.dk

Computational issues

Computational issues

Recall the definition of the distance between two tree-shapes:

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



$\bar{d}(\bar{x}, \bar{w}) = d_1 + d_2 + d_3$

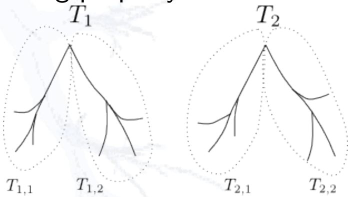
This suggests having to consider infinitely many combinations of paths between different equivalence classes of tree-shapes.

Computational issues

- Recall similarity with TED: computation for unordered trees is NP complete.

Computational issues

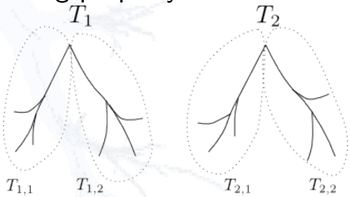
- Recall similarity with TED: computation for unordered trees is NP complete.
- TED has the following property:



$$d(T_1, T_2) = d(T_{1,1}, T_{2,1}) + d(T_{1,2}, T_{2,2}).$$

Computational issues

- Recall similarity with TED: computation for unordered trees is NP complete.
- TED has the following property:



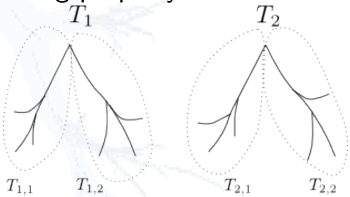
$$d(T_1, T_2) = d(T_{1,1}, T_{2,1}) + d(T_{1,2}, T_{2,2}).$$

- This property is used in many TED algorithms.

madsn@diku.dk

Computational issues

- Recall similarity with TED: computation for unordered trees is NP complete.
- TED has the following property:



  $d(T_1, T_2) = d(T_{1,1}, T_{2,1}) + d(T_{1,2}, T_{2,2})$.
- This property is used in many TED algorithms.
- **The same property does not hold for the QED metric.**

Computational issues

- As a consequence, we expect exact computation of the QED metric to be NP-complete. At present, we approximate by bounding the number k in the definition of the metric:

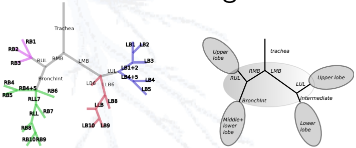
$$\bar{d}(\bar{x}, \bar{y}) = \inf_{k \leq K} \left\{ \sum_{i=1}^{k} d(x_i, y_i) | x_1 \in \bar{x}, y_i \sim x_{i+1}, y_k \in \bar{y} \right\}. \quad (1)$$

Computational issues

- As a consequence, we expect exact computation of the QED metric to be NP-complete. At present, we approximate by bounding the number k in the definition of the metric:

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

- We also reduce computation time using partial labelings, e.g. from an initial branch matching.

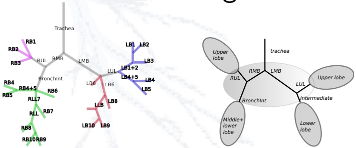


Computational issues

- As a consequence, we expect exact computation of the QED metric to be NP-complete. At present, we approximate by bounding the number k in the definition of the metric:

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

- We also reduce computation time using partial labelings, e.g. from an initial branch matching.



- Finding efficient approximations and heuristics is an extremely important – and interesting – problem!

madsn@diku.dk

Conclusion

- We build a shape space framework for attributed trees.

madsn@diku.dk

Conclusion

- We build a shape space framework for attributed trees.
- We study two natural metrics on the shape space – one, which turns out to be TED, and one, which is new – called QED.

Conclusion

- We build a shape space framework for attributed trees.
- We study two natural metrics on the shape space – one, which turns out to be TED, and one, which is new – called QED.
- Using geometry we explain why TED is not great for doing statistics.

Conclusion

- We build a shape space framework for attributed trees.
- We study two natural metrics on the shape space – one, which turns out to be TED, and one, which is new – called QED.
- Using geometry we explain why TED is not great for doing statistics.
- We show that QED has the types of properties that we need for doing statistics – for instance, we can compute the world's most well-defined average trees!

Conclusion

- We build a shape space framework for attributed trees.
- We study two natural metrics on the shape space – one, which turns out to be TED, and one, which is new – called QED.
- Using geometry we explain why TED is not great for doing statistics.
- We show that QED has the types of properties that we need for doing statistics – for instance, we can compute the world's most well-defined average trees!
- We compute average trees for various types of data.

Open questions

- Statistical properties: How to analyze data variation? PCA analogues and so on?

Open questions

- Statistical properties: How to analyze data variation? PCA analogues and so on?
- Can we change the edgewise distance function?

Open questions

- Statistical properties: How to analyze data variation? PCA analogues and so on?
- Can we change the edgewise distance function?
- Can we generalize to graphs?

Open questions

- Statistical properties: How to analyze data variation? PCA analogues and so on?
- Can we change the edgewise distance function?
- Can we generalize to graphs?
- Can we find efficient algorithms

madsn@diku.dk

Open questions

- Statistical properties: How to analyze data variation? PCA analogues and so on?
- Can we change the edgewise distance function?
- Can we generalize to graphs?
- Can we find efficient algorithms
- Large-scale statistical studies on medical data
  - Geometry-based biomarkers for disease (COPD)?
  - Anatomical modeling?