The Weighted Euler Curve Transform for Shape and Image Analysis

Qitong Jiang
The Ohio State University
Department of Mathematics
jiang.927@osu.edu

Sebastian Kurtek
The Ohio State University
Department of Statistics
kurtek.1@stat.osu.edu

Tom Needham
Florida State University
Department of Mathematics
tneedham@fsu.edu

Abstract

The Euler Curve Transform (ECT) of Turner et al. is a complete invariant of an embedded simplicial complex, which is amenable to statistical analysis. We generalize the ECT to provide a similarly convenient representation for weighted simplicial complexes, objects which arise naturally, for example, in certain medical imaging applications. We leverage work of Ghrist et al. on Euler integral calculus to prove that this invariant—dubbed the Weighted Euler Curve Transform (WECT)—is also complete. We explain how to transform a segmented region of interest in a grayscale image into a weighted simplicial complex and then into a WECT representation. This WECT representation is applied to study Glioblastoma Multiforme brain tumor shape and texture data. We show that the WECT representation is effective at clustering tumors based on qualitative shape and texture features and that this clustering correlates with patient survival time.

1. Introduction

Tools from algebraic topology have become increasingly popular in shape analysis applications over the past several years. At an intuitive level, the topological perspective is appealing because algebraic topology is, at its core, designed to extract tractable algebraic invariants from complex shape data. The dominant technique in topological shape analysis is *persistent homology*, which summarizes multiscale topological features of a shape, where scale is measured relative to some *filtration function*. Roughly, for a continuous function $f : X \to \mathbb{R}$ on a topological space X (satisfying certain tameness conditions), one computes the degree-k homology of the sublevel sets $f^{-1}((-\infty, r])$ and tracks "births" and "deaths" of homological features as the filtration value r is increased. This produces a summary statistic for the pair (X, f) called a *persistence diagram* (see standard references [19, 9]), which can be used as a proxy for X in shape analysis applications. This approach has been taken in several shape analysis tasks, with shape data coming from cortical surfaces [13], brain artery systems [3], proteins [29] and leaf contours [37]. While the persistence diagram of a pair (X, f) provides a computationally tractable shape summary, the complex structure of the invariant means that it is difficult to incorporate into statistical models. A simpler invariant is the *Euler curve* of (X, f); this is an integer-valued function on $\mathbb{R}$ whose value at r is the Euler characteristic (i.e., the alternating sum of ranks of the homology groups) of the sublevel set $f^{-1}((-\infty, r])$.

Given shape data, one must answer the question of which filtration function to apply in order to apply these topological methods. For a shape represented as a simplicial complex K embedded in a Euclidean space $\mathbb{R}^d$, recent work has advocated for using an ensemble of filtration functions given by the height function along directions sampled from the unit sphere S^{d-1} [41, 24, 20, 17, 4, 14, 21]. The collection of all persistence diagrams for these height filtrations is referred to as the persistent homology transform of K. Likewise, the collection of Euler curves for all filtration directions is called the Euler curve transform (ECT) for K. The ECT provides a particularly attractive shape representation, as its simplistic structure allows it to be easily incorporated into statistical models. This was the approach taken in [14], where the ECTs for Glioblastoma Multiforme (GBM) brain tumor shapes were used as covariates in a model for survival prediction.

In this paper, we consider a variant of the ECT, which we dub the *weighted Euler Characteristic Transform* (WECT). This object is defined for shape data consisting of an embedded simplicial complex K endowed with an extra weighting function g. The pair (K, g) is referred to as a *weighted simplicial complex*. The WECT invariant incorporates both the shape of K and the weighting function g into a topological summary. Our motivation for defining this summary also comes from analysis of brain tumor data, which is naturally given as a segmented grayscale image. The segmented shape is used to construct a simplicial complex K embedded in $\mathbb{R}^2$, and the grayscale pixel values inside the shape define the weight function g. While the WECT is a simple

generalization of the ECT, it is able to efficiently incorporate vital information that is ignored by the ECT.

1.1. Contributions and Organization of Paper

The proposed mathematical framework is laid out in detail in Section 2. There, we give a precise definition of the WECT as a generalization of ideas appearing in [41, 4]. We show that recent work of Ghrist, Levanger and Mai implies that the WECT is a complete descriptor of weighted simplicial complexes, i.e., two weighted simplicial complexes have the same WECT if and only if they are equal. In this section, we also provide comparisons between the WECT and other techniques appearing in the topological shape analysis literature. In Section 3, we demonstrate some applications of the WECT framework. We begin with a toy example exploring the utility of the WECT in classifying and registering MNIST digit images. Next, we explore a real application wherein we study the shape and appearance of Glioblastoma Multiforme tumors using WECT representations. Using a simple distance-based clustering scheme, we are able to distinguish clusters of tumors with low survival times, purely from imaging data. Open source code for producing and analyzing WECTs has been made publicly available [27].

2. Mathematical Framework

In this section, we lay out the mathematical framework for the WECT. We begin by reviewing some basic definitions in order to set notation.

2.1. Simplicial Complexes and the Euler Characteristic

Let K be a *simplicial complex* embedded in some Euclidean space $\mathbb{R}^d$. That is, K is a set of embedded *simplices* σ. Each σ is the convex hull of a set of $k+1$ points in general position in $\mathbb{R}^d$, where $k \leq d$ is the *dimension* of the simplex; we write $k = \dim(\sigma)$. For example, a 0-dimensional simplex is a point, a 1-dimensional simplex is a closed line segment and a 2-dimensional simplex is a triangle. The k points defining σ are called its *vertices*. The convex hull of $\ell < k$ of these vertices is also a simplex of K and is called an *ℓ-dimensional face* of σ. If τ is a face of σ, we write $\tau < \sigma$. If σ and τ are simplices of K, we require that $\sigma \cap \tau$ is also a simplex of K. The maximum dimension of a simplex in K is called the *dimension* of K, denoted $\dim(K)$. A collection of simplices of K which itself forms a simplicial complex is called a *subcomplex* of K. The union of all simplices of K of dimension less than or equal to ℓ is a subcomplex called the *ℓ-skeleton* of K, denoted $K^{\leq \ell}$. The set of simplices of K of dimension exactly ℓ is denoted K^ℓ; note that K^ℓ is not a simplicial complex in general.

Abusing notation, we will alternate between treating each embedded simplicial complex as a combinatorial object (a set of simplices) and as a geometric object (a set of points in $\mathbb{R}^d$). We hope that the interpretation should always be clear from context.

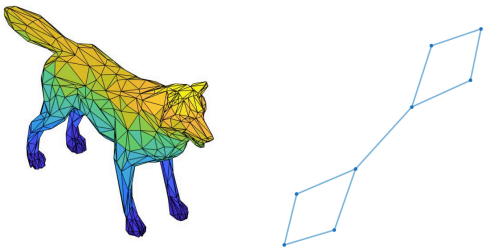

Figure 1. Examples of embedded simplicial complexes commonly arising in computer vision. A triangulated surface is a two-dimensional simplicial complex embedded in $\mathbb{R}^3$. An embedded planar graph is a 1-dimensional simplicial complex in $\mathbb{R}^2$.

A simple combinatorial invariant of a simplicial complex is its *Euler characteristic*, denoted $\chi(K)$. The Euler characteristic is defined as

$$\chi(K) = \sum_{d=0}^{\dim(K)} (-1)^d \cdot \#K^d,$$

where $\#A$ will generally be used to denote the cardinality of a set A. The concept of the Euler characteristic generalizes to more flexible classes of spaces, and it is a basic fact of algebraic topology that χ is a homotopy equivalence invariant. Simplicial complexes form a convenient category for computation, since they can be represented abstractly in a purely combinatorial way by keeping track of all simplices and their inclusions. In this paper, we are focused on the geometrically motivated case where are simplicial complexes are specified by an embedding into a Euclidean space. While not strictly necessary, the invariants we describe are most interesting when $K \subset \mathbb{R}^d$ is a d-dimensional simplicial complex. Moreover, we restrict our attention to the finite setting, i.e., $\#K^\ell$ is finite for all ℓ.

2.2. Euler Curve Transform

Consider a function $f : K \to \mathbb{R}$ as an assignment of a real number to each simplex of K, i.e., the function is constant along faces. The function is a *filtration function* if each sublevel set $f^{-1}((-\infty, r])$ is a subcomplex of K. A filtration function induces a chain of inclusions of simplicial complexes $f^{-1}((-\infty, r_1]) \subset f^{-1}((-\infty, r_2]) \subset \cdots \subset f^{-1}((-\infty, r_n])$, where $r_1 < r_2 < \cdots < r_n$ are the finitely many (using the assumption that K is finite) values in the range of f. From this data, one obtains the *Euler curve* $\chi_f : \mathbb{R} \to \mathbb{Z}$ defined as $\chi_f(r) = \chi\left(f^{-1}((-\infty, r])\right)$.

Given data consisting of an embedded simplicial complex and a relevant function (or more general space and function where similar concepts can be defined), the Eu-

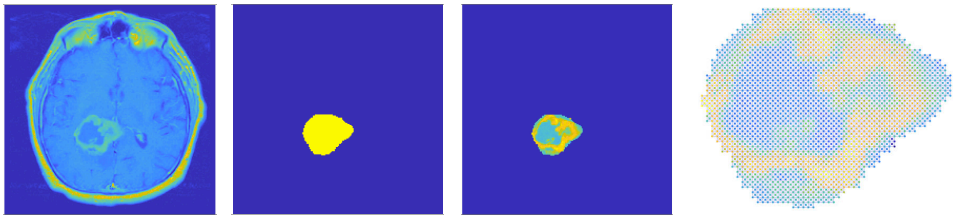

Figure 2. Glioblastoma multiforme tumor image data. From left to right: axial slice with largest tumor area selected from a 3D MRI image; binary tumor segmentation mask; segmented tumor image; weighted simplicial complex created from the segmented tumor image. Observe that the tumor shape data from the segmentation mask is enriched by the overlaid pixel value function extracted from the original image: the level sets of the pixel value function have interesting shape and topological features.

ler curve produces a multiscale topological summary which is amenable to classical analysis, and can be viewed as a simplification of the richer but more computationally taxing persistence diagram [19, 9]. On the other hand, if a relevant function is not provided, one is left with the question of how to filter the simplicial complex.

It was observed in [41] that for an embedded complex $K \subset \mathbb{R}^d$, there is a family of natural filtration functions: orthogonal projections onto the oriented one-dimensional subspaces of $\mathbb{R}^d$, which can be parameterized by the unit sphere $S^{d-1} \subset \mathbb{R}^d$. The *Euler Curve Transform (ECT)* of an embedded simplicial complex $K \subset \mathbb{R}^d$ is the function $\mathrm{ECT}_K : S^{d-1} \times \mathbb{R} \to \mathbb{Z}$ defined as

$$\mathrm{ECT}_K(v, r) = \chi_{p_v}(r),$$

with $p_v : K \to \mathbb{R}$ defined on the vertex set K^0 by the dot product

$$p_v(\sigma) = v \cdot \sigma. \tag{1}$$

The function is extended inductively to higher-dimensional simplices as

$$p_v(\sigma) = \max\{p_v(\tau) \mid \tau < \sigma\}. \tag{2}$$

In practical computations, one uses an approximation of the ECT given by sampling finitely many projection directions from S^{d-1} and finitely many filtration values from $\mathbb{R}$.

One can also apply a smoothing operator to each single variable function $\mathrm{ECT}_K(v, \cdot)$ to obtain the *Smooth Euler Curve Transform (SECT)*. The SECT was applied in [14] to study Glioblastoma Multiforme tumor imaging data. In particular, the SECT served as a shape covariate in a Gaussian process regression model for survival prediction. Another variant of the ECT—very closely related to the one that we consider in subsequent sections—was applied in [4] to provide a topological signature for grayscale image data.

2.3. Weighted Euler Characteristic

Next, suppose that our data consists of an embedded simplicial complex $K \subset \mathbb{R}^d$ together with a function $g : K \to \mathbb{N}$, where $\mathbb{N} = \{1, 2, \ldots\}$. We refer to the pair (K, g) as a *weighted simplicial complex*. The goal is to define a variant of ECT_K, which also incorporates data from g. We note that weighted simplicial complexes have already appeared in the literature in various contexts. To the best of our knowledge, they were first studied in [18], where a homology theory was developed. Abstract weighted simplicial complexes, i.e., those which do not come with a preferred embedding into a Euclidean space, serve as models for collaboration networks [11] and Vietoris-Rips complexes for weighted point clouds [38]. We provide some examples of *embedded* weighted simplicial complexes next.

Example 1. *Our main motivating example comes from grayscale images containing a region of interest, e.g., a tumor image with a segmentation mask, which can be converted into weighted simplicial complexes using Algorithm 1. An example of this process is described in Figure 2.*

Example 2. *Although the main examples considered in this paper will be of the form described in Example 1, we note that there are many other situations where one might wish to consider weighted simplicial complexes. Given shape data as a simplicial complex K, one could consider the weight function g as an annotation or measure of importance. For example, if K is a complex representing a molecule shape, the weight function could be used to annotate different atom types. If K is an anatomical surface, g can be used to indicate regions of importance landmarked by a radiologist.*

For a simplicial complex K and a function $g : K \to \mathbb{N}$, we define the *weighted Euler characteristic*

$$\chi^w(K, g) = \sum_{d=0}^{\dim(K)} (-1)^d \sum_{\sigma \in K^d} g(\sigma).$$

Remark 1. *If $g(\sigma) = 1$ for all $\sigma \in K$, then $\chi^w(K, g) = \chi(K)$. The weighted Euler characteristic is therefore a direct generalization of the classical version.*

Remark 2. *The same definition essentially appears in [4]; the only difference is that only simplicial complexes which are finite axis-aligned lattices were considered there.*

Algorithm 1 Grayscale Image to Weighted Complex

function IMAGETOWEIGHTEDCOMPLEX(A)
  ▷ $A \in \mathbb{N}^{m \times n}$ greyscale image matrix
  $V_{center} = \text{FIND}(A \neq 0)$
  ▷ treat nonzero pixels as coords for vertices
  $V = V_{center}$ ▷ initialize vertex list
  for $v \in V_{center}$ **do** ▷ add corner vertices
    append $v + [\pm 1/2, \pm 1/2]$ to V
  end for
  $V = \text{UNIQUE}(V)$ ▷ remove duplicates
  $F = []$ ▷ initialize face list
  for $v \in V_{center}$ **do**
    append triangles containing v to F
  end for
  $E =$ all resulting edges
  for $f \in F$ containing $v \in V_{center}$ **do**
    $Fw(f) =$ weight of corresponding pixel value
  end for
  for $v \in V$ **do**
    $Vw(v) =$ largest weight of face containing v
  end for
  for $e \in E$ **do**
    $Ew(e) =$ largest weight of face containing e
  end for
  return V, E, F, Vw, Ew, Fw
end function

Remark 3. *A generalization of the weighted Euler characteristic is a classical object of study in algebraic geometry; see, e.g., [28].*

We are particularly interested in functions $g : K \to \mathbb{N}$ which satisfy the consistency condition $g(\tau) = \max\{g(\sigma) \mid \tau < \sigma\}$. Note that this condition is satisfied by the construction given in Algorithm 1. If a function satisfies this condition, we say that it is *admissible*. For functions of this type, the weighted Euler characteristic has a natural interpretation.

Proposition 1. *Suppose that $g : K \to \mathbb{N}$ is an admissible function. Then, each superlevel set $g^{-1}([z, \infty))$ is a subcomplex of K. The weighted Euler characteristic $\chi(K, g)$ is the sum of Euler characteristics of all superlevel complexes of g; that is,*

$$\chi^w(K, g) = \sum_{z \in \mathbb{N}} \chi(g^{-1}([z, \infty))). \tag{3}$$

Proof. We first show that the superlevel sets are subcomplexes of K. It suffices to show that for any $\sigma \in g^{-1}([z, \infty))$ and $\tau < \sigma$, we have $\tau \in g^{-1}([z, \infty))$. This is easy to see from the definition of an admissible function, since $\tau < \sigma$ implies $g(\tau) \geq g(\sigma) \geq z$, which implies $\tau \in g^{-1}([z, \infty))$. It remains to show that Equation (3) is true. In what follows, for a logical statement S, let $\mathbf{1}_S$ denote the indicator function taking the value 1 if S is true, and 0 if S is false. Then,

$$\begin{aligned}
&\sum_{z \in \mathbb{N}} \chi(g^{-1}([z, \infty))) \\
&\quad = \sum_{z \in \mathbb{N}} \sum_{d=0}^{\dim(K)} (-1)^d \#\{\sigma \in K \mid g(\sigma) \geq z\}^d \\
&\quad = \sum_{d=0}^{\dim(K)} (-1)^d \sum_{z \in \mathbb{N}} \#\{\sigma \in K^d \mid g(\sigma) \geq z\} \\
&\quad = \sum_{d=0}^{\dim(K)} (-1)^d \sum_{z \in \mathbb{N}} \sum_{\sigma \in K^d} \mathbf{1}_{g(\sigma) \geq z} \\
&\quad = \sum_{d=0}^{\dim(K)} (-1)^d \sum_{\sigma \in K^d} g(\sigma) = \chi^w(K, g).
\end{aligned}$$

□

2.4. Weighted Euler Curve Transform

We now define the *Weighted Euler Curve Transform (WECT)* as a straightforward generalization of the ECT; the WECT is specifically designed to treat weighted simplicial complexes. Let (K, g) be a weighted simplicial complex, and let $f : K \to \mathbb{R}$ be a filtration function. The *weighted Euler curve* associated to f is the function $\chi^w_f : \mathbb{R} \to \mathbb{Z}$ defined as

$$\chi^w_f(r) = \chi^w(f^{-1}((-\infty, r]), g),$$

where g is understood by context to be the restriction of g to the subcomplex $f^{-1}((-\infty, r])$. We then define the WECT of a weighted simplicial complex (K, g) with $K \subset \mathbb{R}^d$ as the function $\text{WECT}_{K,g} : S^{d-1} \times \mathbb{R} \to \mathbb{Z}$ defined as

$$\text{WECT}_{K,g}(v, r) = \chi^w_{p_v}(r),$$

with p_v the projection function as defined in Equations (1) and (2). Clearly, if the weight function g is constant and equal to one, then $\text{WECT}_{K,g} = \text{ECT}_K$.

As in the case of the ECT, a WECT is represented in practice by sampling a finite number of directions on the sphere S^{d-1}. An example of a WECT is shown in Figure 3. As in [14], when analyzing WECTs, we often preprocess them to improve robustness, by applying a smoothing operator. Unlike [14], we do not specify a particular smoothing operation, and leave the particular method as a hyperparameter in the data analysis pipeline.

2.5. Distance Between WECTs

The WECT of a weighted simplicial complex (K, g) in $\mathbb{R}^d$ is naturally viewed as a family of integer-valued functions $\text{WECT}_{K,g}(v, \cdot) : \mathbb{R} \to \mathbb{Z}$, parameterized by S^{d-1}.

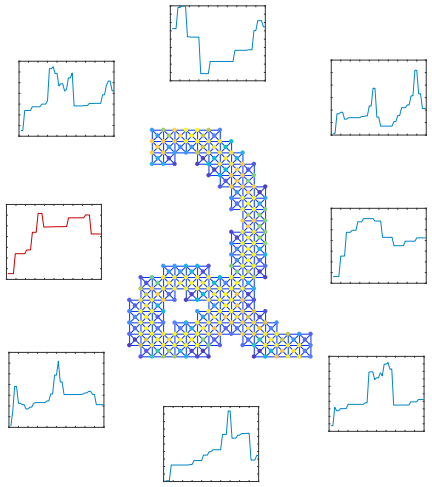

Figure 3. The WECT for a weighted simplicial complex constructed from an MNIST digit. Each panel shows a single weighed Euler curve, with the red curve on the left representing filtering by projection to the vector $(-1, 0)$, and the other curves constructed similarly by projection onto other directions.

Since K is assumed to be compact, each function is constant outside of a compact subset of $\mathbb{R}$, and we may restrict each function to this common compact domain; moreover, given a dataset of weighted simplicial complexes, one may assume without loss of generality that all WECT functions are defined on the same compact domain. After applying a smoothing operator, the smoothed WECT is likewise identified with a parameterized family of compactly supported functions of higher regularity. Any metric d on such functional data gives rise to a metric on WECT data, by integrating the function

$$v \mapsto d(\mathrm{WECT}_{K_1,g_1}(v,\cdot), \mathrm{WECT}_{K_2,g_2}(v,\cdot))$$

over $v \in S^{d-1}$ with respect to its standard volume form.

The most convenient metric on compactly supported functions is the one induced by the standard L^2 norm (with respect to Lebesgue measure), denoted $\|\cdot\|_{L^2}$. We abuse notation slightly and denote the induced metric on the space of WECTs also using norm notation as follows:

$$\|\mathrm{WECT}_{K_1,g_1} - \mathrm{WECT}_{K_2,g_2}\|_{L^2}. \tag{4}$$

This notation is in fact warranted, since this metric is equivalent to the one induced by the L^2 norm on $S^{d-1} \times I$, where I is a compact interval, with respect to the product of the standard measure on S^{d-1} with Lebesgue measure on I. With this metric, the space of WECTs has a Euclidean structure, meaning that WECTs are amenable to methods from functional data analysis and machine learning.

Computationally, a WECT is represented by a finite number of samples. Taking m samples from $\mathbb{R}$ and n samples from S^{d-1}, the values of the WECT can be arranged in a matrix of size $m \times n$. Then, the L^2 distance in Equation (4) can be computed simply as a Frobenius norm, making the process of comparing WECTs numerically efficient.

2.6. Injectivity of the WECT

Inverse problems in topological data analysis have recently become an active topic of research [36]. The basic general question is: Is it possible for inequivalent spaces to be mapped to the same topological summary statistic? This question has recently been tackled for various flavors of topological signatures [22, 35, 15, 12] including Persistent Homology and Euler Curve Transforms [41, 24, 20, 17, 4, 14, 21].

The original paper on the ECT [41] demonstrated a uniqueness result for ECT representations of compact embedded simplicial complexes with an algorithmic proof. This perspective has been pushed further to provide a sufficient number of direction samples to guarantee injectivity [17]. It is shown in [4] that for weighted *cubical complexes* defined on a regular axis-aligned lattice in $\mathbb{R}^d$, only 2^d generic samples are sufficient and an explicit reconstruction algorithm is provided. Our Algorithm 1 produces a simplicial complexes which is essentially equivalent to the cubical complexes of [4], so the reconstruction results their can be ported over directly to weighted simplicial complexes constructed via Algorithm 1.

In anticipation of the possibility of studying non-axis-aligned weighted simplicial complexes through the WECT signature, one might hope for a more general injectivity result. An alternative approach to the injectivity question for ECTs is given in [24, 17]. In these articles, the theory of Euler integral calculus is employed to prove injectivity. This approach is more theoretical and comes with the cost of a less explicit inversion algorithm. This is balanced by more general applicability. In particular, one has the following, quite general, result.

Theorem 1 (Theorem 1, [24])**.** *The map*

$$\mathcal{R} : \mathrm{CF}_c(\mathbb{R}^d) \to \mathrm{CF}(S^{d-1} \times \mathbb{R})$$

defined by

$$(\mathcal{R}(g))(v,r) = \int_{\mathbb{R}^d} g(x) \cdot \mathbf{1}_{x \cdot v \leq r} \, d\chi(x) \tag{5}$$

is injective.

We use $\mathrm{CF}(\mathbb{R}^d)$ to denote the space of *constructible functions*; these are functions $\mathbb{R}^d \to \mathbb{Z}$ whose level sets satisfy a certain tameness condition, defined nowadays in the technical language of o-minimal set theory [2, 16, 24]. The set $\mathrm{CF}(S^{d-1} \times \mathbb{R})$ is defined similarly. We are restricting to *compactly supported* constructible functions $\mathrm{CF}_c(\mathbb{R}^d)$. This space in particular contains admissible functions defined on embedded simplicial complexes in $\mathbb{R}^d$. The right

side of Equation (5) is defined in terms of Euler integration. Roughly, one treats the Euler characteristic formally as a measure, allowing for integration of sufficiently well-behaved functions. The transform $\mathcal{R}$ can be understood as a topological version of the classical Radon transform used in tomography applications [25]. Theorem 1 is proved by appealing to a general result of Schapira on inverting topological Radon transforms of this type [40]. The authors of [24] observe that if g is the indicator function for an embedded simplicial complex K, then $\mathcal{R}(g)$ is exactly the ECT for K, whence the ECT is injective [24, Corollary 1]. On the other hand, if we consider functions g which are admissible weight functions on embedded simplicial complexes, we obtain the following result as an immediate corollary.

Theorem 2. *The Weighted Euler Characteristic Transform is injective on the space of weighted simplicial complexes. That is, if (K_1, g_1) and (K_2, g_2) are weighted simplicial complexes in $\mathbb{R}^d$ with $\mathrm{WECT}_{K_1,g_1} = \mathrm{WECT}_{K_2,g_2}$, then $(K_1, g_1) = (K_2, g_2)$.*

2.7. Comparison to Other Methods

The WECT provides a topological signature which simultaneously incorporates shape data and non-geometric weight data. In the case of image data, by discretely sampling the domain $S^{d-1} \times \mathbb{R}$ one obtains a discrete signature with a similar memory footprint to the original image. However, we show experimentally that the WECT provides a representation, which is more effective at distinguishing shape features. In this subsection, we compare the WECT representation to other shape descriptors appearing in the topological data analysis literature.

Persistent Homology. The WECT representation has several benefits over the commonly used persistence diagram signature. Foremost, it is a nontrivial task to simultaneously incorporate geometric and non-geometric features into a persistence diagram. One approach is to use a multiparameter filtration of the dataset [23, 10]. The major drawback of such an approach is that multiparameter persistent homology does not in general admit a convenient analogue of the persistence diagram statistics used in classical persistent homology. An alternative approach to incorporating geometric and non-geometric features into persistent cohomology was recently proposed in [8], where an enriched barcode representation is obtained through least squares optimization of persistent cohomology cycle representatives.

The simple WECT representation for weighted simplicial complex data also has the benefit of immediately providing a vectorized topological signature. This allows straightforward usage of WECT summaries as covariates in statistical models—this was the main idea of [14], where the ECT summaries were used as covariates in a Gaussian process regression for prediction of survival times of subjects with Glioblastoma Multiforme brain tumors. This is in stark contrast to analysis using persistence diagrams or barcodes from persistent homology. Indeed, a persistence diagram is an unstructured point cloud in $\mathbb{R}^2$ and care must be taken to vectorize this signature in order to incorporate it into statistical models. There are several extant vectorization methods in the literature, including *persistence landscapes* [6] and *persistence images* [1], as well as more straightforward feature aggregation [3]. Any vectorization of the persistence diagram space necessarily distorts its natural latent geometry, since the canonical metric on persistence diagrams, the bottleneck distance, is non-Euclidean [7].

Variants of the ECT. When studying simplicial complexes arising from grayscale image data, one could imagine other relevant simplicial complexes to which one could apply the standard ECT. Examples include thresholding pixel values in the image and building restricted two-dimensional complexes or using the pixel values to build a three-dimensional simplicial complex. We found these approaches to give unsatisfactory performance on our tumor dataset, although they may be viable approaches for other applications.

3. Applications

3.1. Classification of MNIST Digit Images

To understand the descriptive power of the WECT representation of image data, we first explore its ability to classify images from the ubiquitous MNIST handwritten digit dataset [30]. We use a small subset of 1000 28×28 grayscale images, evenly distributed over 10 digits $0, 1, \ldots, 9$. As a baseline, we treat each image as a vector in $\mathbb{R}^{28\times 28}$ and classify them using Support Vector Machines (SVM) with a linear kernel. Next, we produce WECT representations of all digit images. In this experiment, we discretize $S^1 \times \mathbb{R}$ into a 25×50 grid (i.e., 25 Euler curve directions, 50 points along each curve domain). We also smooth the Euler curves to improve robustness using a Gaussian kernel with window size $0.2 \cdot 50$ (these particular parameters were chosen in a tuning step, but we found that the results are generally insensitive to the parameter choice). We then considered each WECT representation as a vector in $\mathbb{R}^{25\times 50}$ and classified using SVM with a linear kernel. We also produced smoothed ECT representations with similar parameters and ran an SVM classification. The ten-fold cross-validated classification rates from these experiments are displayed in Table 1.

The classification results show that the WECT representation of the digit images is adept at encoding and distinguishing shape features, while having a similar memory footprint to the original image representation. It also outperforms the classification using smoothed ECT representations. We stress that this classification result is, of course, not meant to be competitive with those obtained by deep learning methods. Rather, this simple experiment suggests

Table 1. SVM ten-fold classification performance of vectorized image, ECT and WECT representations for the MNIST digit data.

Representation	Classification Rate
Image $\mathbb{R}^{28\times 28}$	87.84 ± 1.42 %
ECT $\mathbb{R}^{25\times 50}$	89.88 ± 1.66 %
WECT $\mathbb{R}^{25\times 50}$	94.68 ± 1.57 %

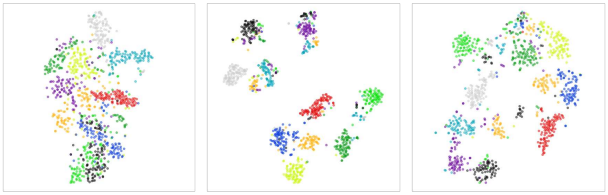

Figure 4. T-SNE embeddings of the MNIST image dataset. Left: Raw image vectors. Middle: Smoothed ECTs. Right: Smoothed WECTs.

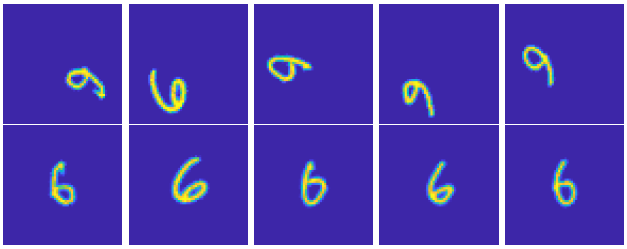

Figure 5. Top: MNIST digit images randomly rotated and translated. Bottom: The same digits after rigid registration to a template digit via the process described in Section 3.2.

that the WECT representation produces an interesting shape summary for this type of image data, which is computationally efficient and can be trivially incorporated into various statistical models.

To get a more detailed qualitative picture of the differences between the raw image, ECT and WECT representations of the MNIST image data, we also computed t-SNE embeddings [31] for each representation; see Figure 4. While class separation is apparent in all three embeddings, it is immediately evident that the embeddings of the ECTs and WECTs are much more distinctly clustered. On the other hand, one can easily see how classification errors arise in the ECT embedding. We believe that these errors occur because the ECT is more sensitive to topological differences between digits, while the WECT smooths these differences using weight data.

3.2. Rigid and Scale Registration

One benefit of the simplicial complex representation of image data is that registering over scale and rigid transformations (translations and rotations) becomes trivial. Once a pair of images have been converted to weighted simplicial complexes (K_1, g_1) and (K_2, g_2), they can be immediately registered with respect to translation and scaling by centering each complex at the origin, and normalizing (treating vertex locations as vectors). To register over rotations, one then computes weighted Euler characteristic transforms WECT_{K_j,g_j}, $j = 1, 2$ and solves the optimization problem

$$\min_{R\in \mathrm{SO}(2)} \|\mathrm{WECT}_{K_1,g_1} - R \cdot \mathrm{WECT}_{K_2,g_2}\|_{L^2}, \qquad (6)$$

where the rotation group $\mathrm{SO}(2)$ acts on a WECT by precomposition in the S^1-coordinate. As was noted above, the L^2 distance is numerically trivial to compute for finite approximations of WECTs. Thus, the optimization problem in Equation (6) can be solved quickly by an exhaustive search over cyclic permutations of the WECT matrix. The minimizing rotation R can then be used to register (K_2, g_2) to (K_1, g_1) with respect to rotations—see Figure 5.

3.3. Analysis of GBM Tumor Data

Glioblastoma Multiforme (GBM) is the most common malignant brain tumor in adults [26]; for most patients, the prognosis is very poor: less than 10% of individuals survive longer than five years and the median survival time is approximately 12 months [42, 34, 33]. GBM is a morphologically heterogeneous disease. GBM tumors exhibit complex structure in terms of their overall shape as well as internal makeup. Often, dead cells are present inside the tumor and increased blood flow near the boundary of the tumor [32]. These features result in various pixel value patterns of GBM tumor images. Thus, characterization of both the shape and texture of GBM tumors, based on medical imaging data, is important for disease prognosis as well as survival prediction. While previous studies have considered these two features separately [5, 39] in the analysis, our approach is to analyze them jointly under a unified representation.

In this study, we use T1-weighted post contrast magnetic resonance images (MRIs) of GBM tumors from 63 subjects. For our analysis, we select a single axial slice with largest tumor area from each 3D image (the same approach was taken in [5, 39]), and summarize the tumors' shapes and textures via the WECT. For details on the image pre-processing steps that were used prior to our analysis, see [39].

We use a simple distance-based clustering approach to analyze the tumor data. First, each of the 63 tumor images is converted into a weighted simplicial complex using Algorithm 1. To isolate the shape and weight information, all simplicial complexes are centered at the origin and normalized so that the vertex farthest from the origin is at distance 1. The weights of the simplicial complexes are then normalized to have maximum weight one; this was done to account for the varying pixel value distributions of the MRIs for each subject. Next, each weighted simplicial complex is given a smoothed WECT representation. Specifically, for each tumor image, we use 25 directions and 50 points along

Table 2. Clusterwise mean and median survival.

Mean	6.7	12.9	20.2
Med.	6.2	9.6	15.2

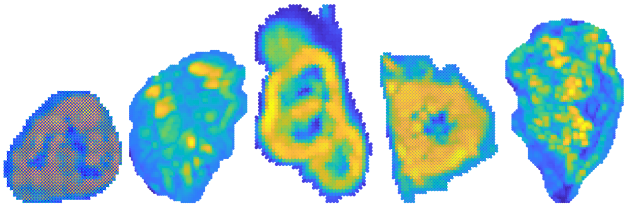

Figure 6. Weighted simplicial complex representations of tumors from the low survival time cluster in Table 2.

Table 3. Clusterwise mean and median survival for Figure 7.

	Blue	Cyan	Red	Magenta	Yellow	Green
Mean	18.1	28.0	17.9	19.4	5.0	12.6
Med.	14.9	22.3	14.3	20.4	4.5	10.7

the domain of the Euler curve for each direction. The Euler curves were smoothed using a Gaussian kernel with a smoothing window of ten. Next, the L^2 distance between each pair of smoothed WECT representations was computed with registration of the tumor images over rotations (see Section 3.2). We applied hierarchical clustering with Ward linkage to the 63×63 distance matrix, which first suggested three natural clusters. The clusterwise mean and median survival times (in months) are reported in Table 2.

These statistics suggest that the clusters are roughly characterized as low, medium and high survival. Figure 6 shows tumors from the low survival cluster; they are visually irregular in shape and intensity distribution, which explains their presence as a distinct cluster. To explore the data in more depth, we consider the clustering dendrogram with this cluster of tumors removed. Figure 7 shows this dendrogram on the remaining 58 tumors, with six highlighted clusters; mean and median survival times for patients in these clusters are shown in Table 3. Inspecting the tumors in these clusters, one can observe various common qualitative shape and intensity features. For example, the tumors in the blue and cyan clusters both tend to have intensity patterns with a ring-like structure near the boundary. The tumors in the blue cluster tend to have higher irregularity in shape and/or intensity patterns, see Figure 8.

4. Future Work

Our work suggests several directions for future research. Driven by the qualitative distance-based clustering results presented here, we next plan to incorporate WECT representations into more sophisticated statistical models for tumor survival prediction. The WECT representation is flexible in the sense that it provides a summary of any weighted simplicial complex. We plan to apply this type of analysis to other shape data, such as weighted simplicial complexes representing annotated molecule shapes. On the theoretical side, there are several interesting questions left open. Principally, one could attempt to strengthen Theorem 2 on injectivity of the WECT in several ways. In its current form, it is mainly a theoretical result and an implementation of an inversion algorithm would be desirable. A practical version of such a construction would only require information about weighted Euler curve measurements in finitely many directions, along the lines of results in [17] on the ECT. It would also be interesting to have a quantitative version of the injectivity theorem; if WECTs of (K_1, g_1) and (K_2, g_2) are close in L^2 distance, does this imply that (K_1, g_1) and (K_2, g_2) are close in some resonable metric, such as Wasserstein distance (treating a normalization of g_j as a probability measure supported on K_j)?

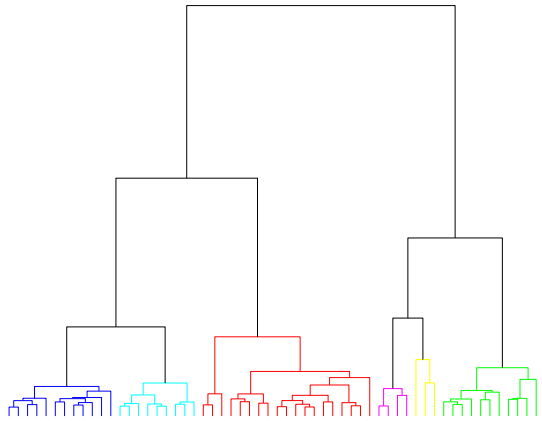

Figure 7. Clustering dendrogram for the tumor dataset with low survival cluster tumors removed.

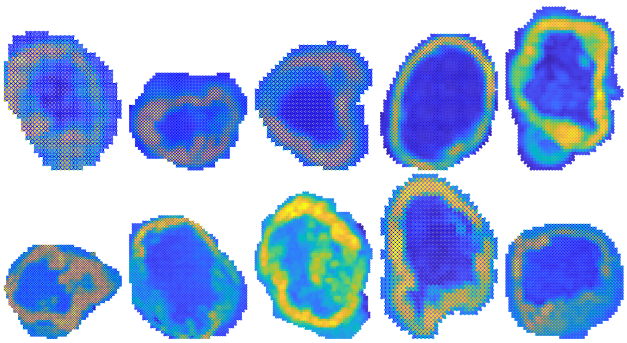

Figure 8. Samples of weighted simplicial complex representations of tumors from cyan (top) and blue (bottom) clusters of Figure 7.

Acknowledgments: We thank Arvind Rao for sharing the GBM dataset. SK was partially supported by NSF DMS-1613054, NSF CCF-1740761, NSF CCF-1839252 and NIH R37-CA214955.

References

[1] Henry Adams, Tegan Emerson, Michael Kirby, Rachel Neville, Chris Peterson, Patrick Shipman, Sofya Chepushtanova, Eric Hanson, Francis Motta, and Lori Ziegelmeier. Persistence images: A stable vector representation of persistent homology. *The Journal of Machine Learning Research*, 18(1):218–252, 2017. 6

[2] Yuliy Baryshnikov, Robert Ghrist, and David Lipsky. Inversion of Euler integral transforms with applications to sensor data. *Inverse problems*, 27(12):124001, 2011. 5

[3] Paul Bendich, James S Marron, Ezra Miller, Alex Pieloch, and Sean Skwerer. Persistent homology analysis of brain artery trees. *The annals of applied statistics*, 10(1):198, 2016. 1, 6

[4] Leo M Betthauser. *Topological Reconstruction of Grayscale Images*. PhD thesis, University of Florida, 2018. 1, 2, 3, 5

[5] Karthik Bharath, Sebastian Kurtek, Arvind Rao, and Veerabhadran Baladandayuthapani. Radiologic image-based statistical shape analysis of brain tumours. *Journal of the Royal Statistical Society, Series C*, 67(5):1357–1378, 2018. 7

[6] Peter Bubenik. Statistical topological data analysis using persistence landscapes. *The Journal of Machine Learning Research*, 16(1):77–102, 2015. 6

[7] Peter Bubenik and Alexander Wagner. Embeddings of persistence diagrams into Hilbert spaces. *arXiv preprint arXiv:1905.05604*, 2019. 6

[8] Zixuan Cang and Guo-Wei Wei. Persistent cohomology for data with multicomponent heterogeneous information. *arXiv preprint arXiv:1807.11120*, 2018. 6

[9] Gunnar Carlsson. Topological pattern recognition for point cloud data. *Acta Numerica*, 23:289–368, 2014. 1, 3

[10] Gunnar Carlsson and Afra Zomorodian. The theory of multidimensional persistence. *Discrete & Computational Geometry*, 42(1):71–93, 2009. 6

[11] Corrie J Carstens and Kathy J Horadam. Persistent homology of collaboration networks. *Mathematical problems in engineering*, 2013, 2013. 3

[12] Michael J Catanzaro, Justin Curry, Brittany Terese Fasy, Jānis Lazovskis, Greg Malen, Hans Riess, Bei Wang, and Matthew Zabka. Moduli spaces of morse functions for persistence. *arXiv preprint arXiv:1909.10623*, 2019. 5

[13] Moo K Chung, Peter Bubenik, and Peter T Kim. Persistence diagrams of cortical surface data. In *International Conference on Information Processing in Medical Imaging*, pages 386–397. Springer, 2009. 1

[14] Lorin Crawford, Anthea Monod, Andrew X Chen, Sayan Mukherjee, and Raúl Rabadán. Predicting clinical outcomes in glioblastoma: an application of topological and functional data analysis. *Journal of the American Statistical Association*, pages 1–12, 2019. 1, 3, 4, 5, 6

[15] Justin Curry. The fiber of the persistence map for functions on the interval. *Journal of Applied and Computational Topology*, 2(3-4):301–321, 2018. 5

[16] Justin Curry, Robert Ghrist, and Michael Robinson. Euler calculus with applications to signals and sensing. In *Proceedings of Symposia in Applied Mathematics*, volume 70, pages 75–146, 2012. 5

[17] Justin Curry, Sayan Mukherjee, and Katharine Turner. How many directions determine a shape and other sufficiency results for two topological transforms. *arXiv preprint arXiv:1805.09782*, 2018. 1, 5, 8

[18] Robert J MacG Dawson. Homology of weighted simplicial complexes. *Cahiers de Topologie et Géométrie Différentielle Catégoriques*, 31(3):229–243, 1990. 3

[19] Herbert Edelsbrunner and John Harer. *Computational topology: an introduction*. American Mathematical Soc., 2010. 1, 3

[20] Brittany Terese Fasy, Samuel Micka, David L Millman, Anna Schenfisch, and Lucia Williams. Challenges in reconstructing shapes from Euler characteristic curves. *arXiv preprint arXiv:1811.11337*, 2018. 1, 5

[21] Brittany Terese Fasy, Samuel Micka, David L Millman, Anna Schenfisch, and Lucia Williams. Persistence diagrams for efficient simplicial complex reconstruction. *arXiv preprint arXiv:1912.12759*, 2019. 1, 5

[22] Patrizio Frosini and Claudia Landi. Uniqueness of models in persistent homology: the case of curves. *Inverse problems*, 27(12):124005, 2011. 5

[23] Patrizio Frosini, Michele Mulazzani, et al. Size homotopy groups for computation of natural size distances. *Bulletin of the Belgian Mathematical Society-Simon Stevin*, 6(3):455–464, 1999. 6

[24] Robert Ghrist, Rachel Levanger, and Huy Mai. Persistent homology and Euler integral transforms. *Journal of Applied and Computational Topology*, 2(1-2):55–60, 2018. 1, 5, 6

[25] Sigurdur Helgason and S Helgason. *The radon transform*, volume 2. Springer, 1980. 6

[26] Eric C Holland. Glioblastoma multiforme: The terminator. In *Proceedings of the National Academy of Sciences*, volume 97, pages 6242–6244, 2000. 7

[27] Qitong Jiang, Sebastian Kurtek, and Tom Needham. Weighted Euler curve transform github repository. `https://github.com/trneedham/Weighted-Euler-Curve-Transform`. 2

[28] Masaki Kashiwara. Index theorem for constructible sheaves. *Astérisque*, 130:193–209, 1985. 4

[29] Violeta Kovacev-Nikolic, Peter Bubenik, Dragan Nikolić, and Giseon Heo. Using persistent homology and dynamical distances to analyze protein binding. *Statistical applications in genetics and molecular biology*, 15(1):19–38, 2016. 1

[30] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998. 6

[31] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(Nov):2579–2605, 2008. 7

[32] Andriy Marusyk, Vanessa Almendro, and Kornelia Polyak. Intra-tumour heterogeneity: A looking glass for cancer? *Nature Reviews Cancer*, 12(5):323–334, 2012. 7

[33] Roger McLendon, Allan Friedman, Darrell Bigner, Erwin G Van Meir, Daniel J Brat, Gena M Mastrogianakis, Jeffrey J Olson, Tom Mikkelsen, Norman Lehman, Ken Aldape, et al. Comprehensive genomic characterization de-

fines human glioblastoma genes and core pathways. *Nature*, 455(7216):1061–1068, 2008. 7

[34] Mairéad G McNamara, Solmaz Sahebjam, and Warren P Mason. Emerging biomarkers in glioblastoma. *Cancers*, 5(3):1103–1119, 2013. 7

[35] Steve Oudot and Elchanan Solomon. Barcode embeddings for metric graphs. *arXiv preprint arXiv:1712.03630*, 2017. 5

[36] Steve Oudot and Elchanan Solomon. Inverse problems in topological persistence: a survey. In *Abel Symposia*, 2019. 5

[37] Vic Patrangenaru, Peter Bubenik, Robert L Paige, and Daniel Osborne. Challenges in topological object data analysis. *Sankhya A*, pages 1–28, 2018. 1

[38] Shiquan Ren, Chengyuan Wu, Jie Wu, et al. Weighted persistent homology. *Rocky Mountain Journal of Mathematics*, 48(8):2661–2687, 2018. 3

[39] A. Saha, S. Banerjee, S. Kurtek, S. Narang, J. Lee, G. Rao, J. Martinez, K. Bharath, A.U.K. Rao, and V. Baladandayuthapani. DEMARCATE: Density-based magnetic resonance image clustering for assessing tumor heterogeneity in cancer. *NeuroImage: Clinical*, 12:132 – 143, 2016. 7

[40] Pierre Schapira. Tomography of constructible functions. In *International Symposium on Applied Algebra, Algebraic Algorithms, and Error-Correcting Codes*, pages 427–435. Springer, 1995. 6

[41] Katharine Turner, Sayan Mukherjee, and Doug M Boyer. Persistent homology transform for modeling shapes and surfaces. *Information and Inference: A Journal of the IMA*, 3(4):310–344, 2014. 1, 2, 3, 5

[42] B Tutt. Glioblastoma cure remains elusive despite treatment advances. *OncoLog*, 56(3):1–8, 2011. 7