Communications for Statistical Applications and Methods
2024, Vol. 31, No. 3, 309–322

https://doi.org/10.29220/CSAM.2024.31.3.309
Print ISSN 2287-7843 / Online ISSN 2383-4757

# Classification of algae in watersheds using elastic shape

Tae-Young Heo[a], Jaehoon Kim[a], Min Ho Cho[1,b]

[a]Department of Information Statistics, Chungbuk National University, Korea;
[b]Department of Statistics, Inha University, Korea

**Abstract**

Identifying algae in water is important for managing algal blooms which have great impact on drinking water supply systems. There have been various microscopic approaches developed for algae classification. Many of them are based on the morphological features of algae. However, there have seldom been mathematical frameworks for comparing the shape of algae, represented as a planar continuous curve obtained from an image. In this work, we describe a recent framework for computing shape distance between two different algae based on the elastic metric and a novel functional representation called the square root velocity function (SRVF). We further introduce statistical procedures for multiple shapes of algae including computing the sample mean, the sample covariance, and performing the principal component analysis (PCA). Based on the shape distance, we classify six algal species in watersheds experiencing algal blooms, including three cyanobacteria (Microcystis, Oscillatoria, and Anabaena), two diatoms (Fragilaria and Synedra), and one green algae (Pediastrum). We provide and compare the classification performance of various distance-based and model-based methods. We additionally compare elastic shape distance to non-elastic distance using the nearest neighbor classifiers.

Keywords: algal blooms, shape of algae, elastic metric, square root velocity function, principal component analysis

## 1. Introduction

Shape is an important physical property of an object that characterizes its appearance. It is often represented by the boundary of the object in an image or a video as in Figure 1. Using this representation as data, a variety of statistical analysis has been conducted. Particularly, the classification of shape is one of the most fundamental tasks in many application fields, ranging from medical imaging, and computer vision to bioinformatics. It can also be applied to environmental science and engineering.

Traditionally, various mathematical representations of shape were proposed and developed, which include (unordered) point clouds (Besl and McKay, 1992), a set of (ordered) finite points called landmarks (Dryden and Mardia, 2016), level sets (Malladi *et al.*, 1996), skeletal models (Pizer *et al.*, 2013), and diffeomorphic transforms or deformable templates (Grenander and Miller, 1998). Kendall (1984) defined the shape as the geometric information of an object after filtering out translation, scaling and rotation. Thus, there were extensive works for shape analysis to deal with this invariance property. It was shown that the geometric space of a shape is not Euclidean, and thus an appropriate metric was needed to quantify shape differences on the corresponding space.

For Min Ho Cho, this work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korean government (RS-2022-00167077) and INHA UNIVERSITY Research Grant.

[1]Corresponding author: Department of Statistics, Inha University, 100 Inha-ro, Michuhol-gu, Incheon 22212, Korea. E-mail: mcho@inha.ac.kr

Published 31 May 2024 / journal homepage: http://csam.or.kr

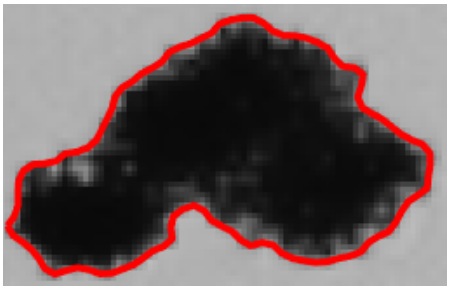

Figure 1: *A boundary of one alga sample of the Microcystis species, represented as a planar closed curve.*

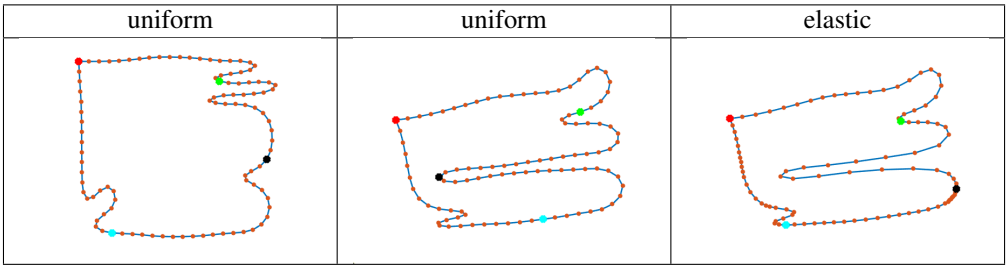


Figure 2: *Two different curves of algae with uniform parameterization (left and middle), and one curve (right) with the same shape as the middle curve but with different parameterization optimally registered to the left curve.*

However, there have been steady efforts to develop a framework for the representation of shape using its entire contour instead of a finite number of points. The continuous curve seems to be more natural when analyzing its shape. Additionally, there is an issue regarding how and where to choose landmarks if we represent shape by discrete points. As a result, functional representations of curves for shape have recently been developed. For instance, the shape of an alga in Figure 1 is represented as a planar curve, so regarded as two-dimensional functional data. Yet, we need to consider the invariance of an additional transformation as well as translation, scaling, and rotation when comparing shapes of two different continuous curves. This is re-parameterization, which is a smooth one-to-one transformation of the domain of curves. For instance, the two curves (middle and right) in Figure 2 have an identical shape no matter how we parameterize the curves.

Moreover, if we assume that the curve is closed such as in Figure 1, we need a closure condition such that the mapping from a domain to the curve describes the traversal of the shape with the same starting and ending points. To solve these challenges, mathematical frameworks under the Riemannian structure have been developed by Michor and Mumford (2006) and Srivastava *et al.* (2007). We adopt a recent framework called the elastic shape analysis framework in this paper. Brief geometrical backgrounds and preliminaries of this framework are described in Section 2.

## 1.1. Motivation

There has been an increasing need for studying algal species in water as the overgrowth of algae has a critical impact on the management of drinking water supply systems (Coltelli *et al.*, 2014). Such massive algal blooms in aquatic ecosystems often trigger undesirable effects on the quality of drinking water, which include algal toxins, an unfavorable odor and taste (Paerl and Otten, 2013). A wide range

of research on algae has been conducted over several decades. For example, experiments to determine the growth rate of some environmental factors, such as temperature and light intensity, have been performed (Dauta *et al.*, 1990; Paerl and Otten, 2013). Understanding differences and similarities between algal species can help in distinguishing species and measuring their health.

Regular monitoring of algal blooms is an important task for ensuring the safety of water supply systems. The direct counting of algal cells using a microscope is a traditional method for monitoring the status of algal blooms, but it is a time-consuming process that requires intensive labor from researchers. Therefore, efforts to develop automated technology to reduce time and effort in algal cell identification have continued. Understanding the differences and similarities between algal shapes can help in the identification of algal genera. The algal cell image dataset used in this study was collected using a digital imaging flow cytometer and microscope (FlowCAM, Fluid Imaging Technologies, Yarmouth, ME, USA) provided by Korea Water Resources Corporation (K-water). The dataset includes a total of 2571 morphological images of six algal genera, including three cyanobacteria (Microcystis sp., Oscillatoria sp., and Anabaena sp.), two diatoms ( Fragilaria sp. and Synedra sp.), and a green alga (Pediastrum sp.) (Park *et al.*, 2019).

Many approaches for morphological identification of algae in watersheds have been developed based on its images, which were captured by some microscope devices such as a digital imaging flow cytometer and microscope (FlowCAM). Using the pixel values with some image analysis tools, many methods are proposed for the classification of algae, including the convolutional neural network (CNN) (Medina *et al.*, 2017). The machine learning analysis for algae images and a novel framework combining the CNN and a neural architecture search (NAS) technologies are proposed (Park *et al.*, 2019).

However, the shape of algal species has seldom been studied although it is one of the most important features for identification. The shape of algal species, which is represented as a planar closed curve that is extracted from its microscope image, is suitable for classification. In this paper, we adopt and describe the elastic shape analysis framework with a novel functional representation called the square-root velocity function (SRVF). The benefits from using this representation and the elastic Riemannian metric are described along with the inherent geometry. Based on this mathematical framework, we can define the shape distance and further shape statistics when multiple sample curves are given. We then apply various well-known statistical classification methods to the dataset of algal shape. One group of the methods is based on pairwise shape distances such as the nearest neighbor classifiers. The other group is based on probability distributions for shape such as linear and quadratic discriminant analysis in the standard multivariate fashion.

## 1.2. Contributions

In this paper, our contributions are as follows: (i) introducing various classification approaches for algae based on the elastic shape analysis framework, (ii) providing the experimental results from real environmental systems, and (iii) comparing and investigating the strengths and drawbacks of the presented approaches.

Under the elastic shape distance with the SRVF representation after removing translation, scaling, rotation and re-parameterization, we first evaluate the classification performance using the algal shape of $k$-nearest-neighbors. We compare the classification result using the non-elastic shape distance that does not allow re-parameterization. In other words, the non-elastic shape distance sticks to fixed parameterization when matching two curves. Not only is there a numerical comparison of the classifiers based on the two distances, but also the difference in their visual deformations between two curves is included. An additional distance-based procedure is considered and compared. It also chooses the

nearest species via their average shape. We call this classifier the nearest mean method. Through the empirical study of algal classification, we provide the accuracy of classification only by algal shape and highlight the value of using the elastic distance.

Since curves reside in non-linear manifold, it is complicated to build a probability model on their representation space. Even though the SRVF representation simplifies the elastic metric to the simple $\mathbb{L}^2$ metric and the corresponding space becomes the unit Hilbert sphere, it is still non-Euclidean. Thus, we utilize a tangent space at a particular point on the sphere and a projection. Analytic expressions of some tools for this projection are well-known in differential geometry. Among many choices of the projection, we use the inverse-exponential map to project the SRVFs on the shape space into a tangent space produced at the sample mean shape. Since it is now a Euclidean space, we can use the standard principal component analysis (PCA) to reduce dimension of the data and construct a probability model with the projected shape representations.

We first apply the Gaussian distribution with low-dimensional principal components on the tangent space and classify some test shapes by the Gaussian likelihoods after estimating the mean and covariance for each species. Assuming both equal and unequal covariance structures as linear and quadratic discriminant analysis (LDA & QDA), we set these model-based classification procedures as baseline (Pal *et al.*, 2017). We further consider LDA- and QDA-type classifiers using the aggregated likelihoods from all possible pairwise PC subspaces (Cho *et al.*, 2021). More details of these methods are described in Section 3.2. By comparing the classification results with various choices of PC dimension, we can see some patterns and optimal dimensions of the classification accuracy for algal shape.

The rest of this paper is organized as follows. Section 2 briefly reviews the geometric framework for elastic shape analysis of planar curves. Section 3 begins by describing various classification approaches. First a few nonparametric methods are based on the pairwise distances, such as the nearest neighbors and the nearest mean rules. The other model-based methods need the computation of some relevant statistics and the estimation of appropriate probability distributions. We then introduce two standard procedures and two additional ones, which rely on pairwise statistics and dimension reduction to different degrees. Section 4 provides empirical studies that show applications of these diverse procedures in shape classification for algal species. Section 5 provides a short discussion and lays out some directions for future work.

## 2. Geometric background

In this section, we briefly describe a Riemannian geometric framework for non-Euclidean space that shape of curves reside in. Among many functional representations and metrics of curves for shape, we adopt the square-root velocity function (SRVF) and the elastic Riemannian metric to compute the shape distance. More details of this elastic shape analysis framework are provided by Srivastava and Klassen (2016).

### 2.1. Functional representation and metric for shape

We represent the shape of an algae as an absolutely continuous, parameterized curve in $\mathbb{R}^2$. For the closed curves that we used, they are denoted as $\beta : \mathbb{S}^1 \to \mathbb{R}^2$ where the domain $\mathbb{S}^1$ is a unit circle which implies that the starting and the ending points of $\beta$ are the same. Since closed curves are handled similarly to open curves with minor adjustments, we describe the framework with open curves $\beta : [0, 1] \to \mathbb{R}^2$. To extract a curve of the algae from an image, we take its outline by the sequence of 2D coordinates $\beta(t) = (x(t), y(t))$. The process of taking the boundary is described in

Section 4.2.

Once we set the functional representation for algal shape, we next need an appropriate distance between two curves $\beta_1$ and $\beta_2$. However, in shape analysis, we must ensure the invariance property of the distance to translation $T \in \mathbb{R}^2$, scaling $s \in \mathbb{R}_+$, rotation $O \in SO(2) = \{O \in \mathbb{R}^{2\times 2} | O^T O = OO^T = I,\ \det(O) = +1\}$ (called the special orthogonal group), and re-parameterization $\gamma \in \Gamma = \{\gamma : [0,1] \to [0,1] | \gamma(0) = 0,\ \gamma(1) = 1,\ \dot{\gamma} > 0,\ \gamma \text{ is a diffeomorphism}\}$, where $\dot{\gamma}$ is the derivative of $\gamma$. The translation $\beta + T$ and the scaling $c\beta$ are relatively easy to filter out of the representation by relocating and normalizing its size. However, for the rotation $O\beta$ and the re-parameterization $\beta \circ \gamma$, it is more challenging to achieve the invariance of these two transformations.

The $\mathbb{L}^2$ distance between two curves given by $\|\beta_1 - \beta_2\| = \sqrt{\int_0^1 |\beta_1(t) - \beta_2(t)|^2 dt}$ seems a natural choice. The norm $|\cdot|$ of the integrand denotes the Euclidean (vector) norm in $\mathbb{R}^2$. However, this distance is not parameterization invariant because $\|\beta_1 - \beta_2\| \neq \|\beta_1 \circ \gamma - \beta_2 \circ \gamma\|$ is for a general re-parameterization $\gamma \in \Gamma$ (Srivastava *et al.*, 2011). Thus, other types of distance between two curves have been considered. Mio *et al.* (2007) defined the elastic metric on the Riemmanian manifold which consists of instantaneous speed and instantaneous direction components at any point. The authors showed that this elastic metric is invariant to re-parameterization as well as the other shape preserving transformations. Nevertheless, the direct use of the elastic metric to $\beta$ was limited due to computational difficulties. To overcome this problem, Joshi *et al.* (2007) and Srivastava *et al.* (2011) introduced a new functional representation for shape called the square-root velocity function (SRVF), given by

$$q(t) \equiv \begin{cases} \dot{\beta}(t)/\sqrt{|\dot{\beta}(t)|}, & \text{if } |\dot{\beta}(t)| \neq 0, \\ 0, & \text{otherwise,} \end{cases} \tag{2.1}$$

where $\dot{\beta}(t)$ is the derivative of $\beta$ at $t$. The SRVF representation has several benefits for shape analysis. First, it simplifies the elastic metric to the $\mathbb{L}^2$ metric, which facilitates the computation of distances between two curves (Srivastava *et al.*, 2011). They showed that the $\mathbb{L}^2$ distance on the space of SRVFs is equivalent to a particular elastic distance on the space of curves. The $\mathbb{L}^2$ distance between SRVFs is invariant to rotation and re-parameterization (Kurtek *et al.*, 2012). Furthermore, since we defined $\beta$ as an absolutely continuous curve, its SRVF $q$ is square-integrable and is referred to simply as $\mathbb{L}^2$ (Robinson, 2012). Finally, one can uniquely recover the curve $\beta$ from its SRVF $q$ up to a translation, using the equation $\beta(t) = \beta(0) + \int_0^t q(s)|q(s)|ds$ where $t = 0$ is the start point of the parameterization.

## 2.2. Shape space and shape distance

Since the SRVF representation $q$ is composed with the derivative of the curve $\beta$, the translation is automatically removed. Once we normalize $q$ to have unit length, the scaling variability of the curve $\beta$ is also removed. Then, the set of all normalized SRVFs, $C = \{q \mid \|q\|^2 = \int_0^1 |q(t)|^2 dt = \int_0^1 |\dot{\beta}(t)| dt = 1\}$ is called the pre-shape space, and it is geometrically shown that it is the unit Hilbert sphere in $\mathbb{L}^2([0,1]], \mathbb{R}^2)$. Thus, the distance between any two different $q_1,\ q_2 \in C$ is an arc length of the great circle of the unit sphere, and is computed by $d_C(q_1, q_2) = \cos^{-1}(\langle q_1, q_2 \rangle)$, where $\langle q_1, q_2 \rangle = \int_0^1 q_1(t)^T q_2(t) dt$.

However, we still need to consider the rotation and re-parameterization of the curves. We use the concept of equivalence class, defined by $[q] = \{O(q, \gamma) | O \in SO(2), \gamma \in \Gamma\}$. Each equivalence class is a unique shape. Then, we can define the shape space as the set of all equivalence classes,

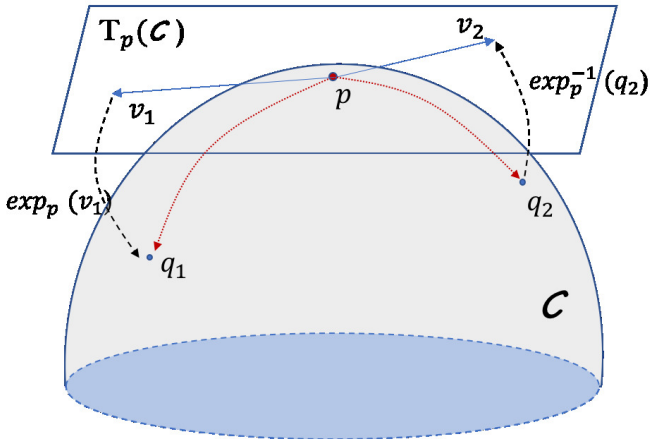


Figure 3: *Two SRVFs $q_1$ and $q_2$ and their corresponding representations $v_1$ and $v_2$ in the tangent space, computed by the inverse-exponential map at a point of projection $p$.*

$\mathcal{S} = C/(SO(2) \times \Gamma) = \{[q] | q \in C\}$. It means that the shape space $\mathcal{S}$ is a quotient space of the pre-shape space $C$ under the action of the rotation and re-parameterization groups. The shape distance, or geodesic between two curves is defined as the distance between their equivalence classes, and is computed by

$$d_{\mathcal{S}}([q_1], [q_2]) = \min_{O \in SO(2), \gamma \in \Gamma} \cos^{-1}(\langle q_1, O(q_2, \gamma) \rangle). \tag{2.2}$$

If we denote the minimizers of Equation (2.2) as $O^*$ and $\gamma^*$, the geodesic computation involves finding $q_2^* = O^*(q_2, \gamma^*) \in [q_2]$ after we fix one $q_1 \in [q_1]$. We call this the registration problem. In other words, when measuring pure shape dissimilarity between two curves, the step of finding the correspondence of points across two curves by allowing elastic parameterization is vital as shown in the right panel in Figure 2. Thus, we compute shape distance by aligning curves in an elastic way as computing distance between functions based on the dynamic time warping. The solution $O^*$ is obtained using singular value decomposition (Srivastava and Klassen, 2016), known as a part of the Procrustes analysis. With $O$ fixed, $\gamma^*$ is computationally obtained by the Dynamic Programming algorithm (Robinson, 2012) which iterates searching over all paths in $\Gamma$ and results in an approximation to $\gamma^*$. We then hold the parameterization fixed and find $O^* \in SO(2)$ again. We reiterate between these two steps until the convergence criterion is satisfied. In the case of closed curves, one must additionally perform an exhaustive search for the optimal starting point on the shape. After registration, one can construct a geodesic path between two shapes by connecting $q_1$ and $q_2^*$ on the pre-shape space $C$.

## 2.3. Shape statistics

Under the space of SRVFs with the distance that is invariant to all shape-preserving transformations, we define sample statistics for the shape of curves, which include the sample mean and the sample covariance. Although we derived the pre-space of the SRVFs as the unit sphere where computations are relatively easy, this transformed space is not Euclidean. Thus, traditional vector calculus cannot be applied. We adopt the concept of the tangent space so that standard statistical procedures are applicable on this linearized space. To approximate the tangent space, we need to choose an appropriate point on the sphere for projection and a tool for moving SRVFs onto the tangent space. We use the

exponential map and its inverse as the projection tool in differential geometry. Figure 3 illustrates these two maps with $C$ via a three-dimensional unit sphere and $T_p(C)$ at $p \in C$. The exponential map ($\exp_p : T_p(C) \to C$) and the inverse exponential map ($\exp_p^{-1} : C \to T_p(C)$) are given by

$$\begin{aligned} \exp_p(v) &= \cos(\|v\|)\, p + \sin(\|v\|)\, \frac{v}{\|v\|} = q, \\ \exp_p^{-1}(q) &= \frac{\theta}{\sin(\theta)}\,(q - \cos(\theta)p) = v, \end{aligned} \tag{2.3}$$

for $p, q \in C,\ v \in T_p(C)$, and where $\|\cdot\|$ is the $\mathbb{L}^2$ norm and $\theta = \cos^{-1}(\langle q_1, q_2 \rangle)$. The exponential map transfers vectors along geodesics from a tangent space to the pre-shape space, and the inverse map moves points from the nonlinear space to the tangent space. The projected vector in the tangent space preserves the same length and direction as those of the geodesic arc from the projection point to its SRVF in the pre-shape space.

For a projection point which determines the tangent space and also affects any result of subsequent statistical analysis, the sample mean is one of the most commonly used. Once we have a sample of curves, we convert the curves into normalized SRVFs $q_1, \ldots, q_n \in C$, and then the sample mean is computed using the shape distance $d_{\mathcal{S}}$ given in Equation (2.2):

$$[\bar{q}] = \arg\min_{[q] \in \mathbb{S}} \sum_{i=1}^{n} d([q], [q_i])^2 . \tag{2.4}$$

It is called the sample Karcher mean. While this mean is an entire equivalence class based on the definition, we proceed with subsequent analysis simply by selecting one element $\bar{q} \in [\bar{q}]$. The computation of the Karcher mean is to solve the optimization problem in Equation (2.4) which involves iterative pairwise alignments with one $q$ fixed and iterative mappings between the curved space and the linearized space. This optimization problem is often solved via a gradient descent approach by Kurtek *et al.* (2013). The detailed algorithm is given in Dryden and Mardia (2016).

Given sample SRVFs and their mean shape, we can define the Karcher covariance on a locally linearized tangent space. Let $v_i = \exp_{\bar{q}}^{-1}(q_i^*) \in T_{[\bar{q}]}(\mathcal{S}),\ i = 1, \ldots, n$ be the $i^{th}$ projected vector after being registered to the mean shape similar to Equation (2.2). Since we use the functional representation, the computed covariance for the vectors $v_i$ is inherently in the infinite dimensional tangent space. However, in practice, the curves are sampled using a finite number of points, say $m$. So, the observed tangent data matrix is formed as $V \in \mathbb{R}^{n \times 2m}$, where a long vector of size $2m$ is made by stacking the $x, y$ coordinates for each $v_i$. Then, we can compute the Karcher covariance matrix, $K \in \mathbb{R}^{2m \times 2m}$, and use $Q = (1/(n-1))V^T V$. Since the sample mean is the projection point, it is the origin of the tangent space, $\bar{V} = 0$.

## 3. Classification approaches of algal shapes

We describe three distance-based methods and four different model-based procedures for the classification of shape data that involves nonlinear registration and resides in non-Euclidean space. The presented classification approaches are all based on the elastic shape analysis framework introduced in the previous section.

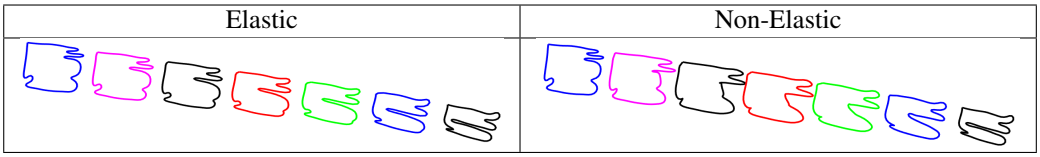


Figure 4: *Comparison of elastic vs. non-elastic shape deformations.*

### 3.1. Distance-based approaches

We first consider two well-known nonparametric classification methods: (1) $k$-nearest neighbors, and (2) the nearest mean classifiers. For the nearest neighbors, we use and compare the elastic and non-elastic distance between two curves. We randomly split data into training and test parts. We compute geodesic or the elastic distance in Equation (2.2) between a test shape and all training shapes in a pairwise manner. Specifically, we fix each test shape $q_i^{te}$ in its equivalence class for $i = 1, \ldots, n_{te}$ and find the optimally registered $q_j^{tr*}$, $j = 1, \ldots, n_{tr}$ for each training shape. Once we compute all pairwise geodesic distances, $d_{\mathcal{S}}([q_i^{te}], [q_j^{tr}])$, we find the $k$ training shapes that have the nearest distance to a given test shape $q_i^{te}$. Assuming the true classes of training shapes are known, we classify the given test shape $q_i^{te}$ into the class of the majority of $k$ neighbors. We could also fix a training shape and then find the optimally registered test shapes. We then can expect identical classification results since the two geodesic distances in both directions are theoretically equivalent.

The non-elastic distance is the minimizer of Equation (2.2) except for the step of finding the optimal re-parameterization. That is, $\min_{O \in SO(2)} \cos^{-1}(\langle q_1, Oq_2 \rangle)$, and the distance with points correspondence between the left and the middle curves in Figure 2 after alignment of rotation. Figure 4 visualizes shape deformations along geodesic between two shapes at each end, and compares the elastic and non-elastic distances. Visually, the elastic geodesic represents more natural deformations between shapes. Important features are preserved along the path. In addition, different classification results are expected between the two distances since these distances are different for the identical pair of shapes.

The nearest mean classifier first finds the Karcher sample mean of the training shapes in each class using Equation (2.4). Then, the shape distance from each test case to the shape mean for each class is computed using Equation (2.2). Finally, the test case is assigned to the class giving the minimum distance. The nearest mean classifier enables us to reduce computational cost compared to $k$-nearest neighbors since it only requires computation of the distance from a test shape to each training class mean and the estimation of the classwise means. However, the classification accuracy of this approach might be inferior because it classifies a test shape using only a single representative shape for each group (the mean shape).

### 3.2. Model-based approaches

Based on the Karcher sample mean and covariance under the elastic shape analysis framework with the SRVF representation, we consider four model-based classification methods by estimating appropriate probability distributions on the tangent space: (1) linear and quadratic discriminant analysis (LDA & QDA) on a single space (Pal *et al.*, 2017), and (2) the LDA- and QDA-typed classifiers on multiple spaces by aggregating likelihoods from pairwise comparisons if there are more than two classes (Cho *et al.*, 2021). The LDA and QDA classification approaches rely on covariance matrices. Yet, typical shape data suffers from the high dimensional problem, for example $n << 2m$, though we

use discretized points for shape representation in practice. Then, the estimated covariance matrices become singular. This necessitates dimension reduction, so the models we present are built on the reduced space spanned by a few tangent principal components (tPCs) (Dryden and Mardia, 2016). First, we use singular value decomposition to compute $Q = \mathrm{UDU}^T$, where $U$ is an orthonormal matrix with columns specifying the principal directions of shape variation, and $D$ is a diagonal matrix with non-negative entries arranged in decreasing order to specify the principal component variances. By selecting $r < n - 1$, one has a lower-dimensional Euclidean representation of the shapes in the tangent space as $Z \in \mathbb{R}^{n\times r}$, with $z_{ij} = v_i U_j,\ i = 1, \ldots, n,\ j = 1, \ldots, r$. These tangent PC data are used for classification of shapes with LDA or QDA.

The first two classifiers use probability models on a single tangent space as a baseline. Let $\bar{q}$ be the overall mean over all classes. We then project all training shapes into a tangent space at $\bar{q}$. The covariance matrix pooled over all $K$ classes is estimated and $r$ dimensional tPC coefficients are obtained. The log-likelihood of a test shape $x$, also projected on the $r$ dimensional tPC subspace, is calculated assuming the multivariate Gaussian as follows.

$$l_{\bar{q}}\left(x; \hat{\mu}_k, \hat{\Sigma}_k\right) = -\frac{1}{2}\log\left|2\pi\hat{\Sigma}_k\right| - \frac{1}{2}\left(x - \hat{\mu}_k\right)^T \hat{\Sigma}_k^{-1}\left(x - \hat{\mu}_k\right), \tag{3.1}$$

where $\hat{\mu}_k$, $\hat{\Sigma}_k$ are the mean vector and the covariance matrix estimated from the tPC coefficients for class $k$. For LDA, we use $\hat{\Sigma}_P = (1/K)\sum_{k=1}^{K}\hat{\Sigma}_k$, which assumes a balanced situation, in place of each $\hat{\Sigma}_k$. Finally, we choose the class with the largest log-likelihood for the test shape $x$.

The next two model-based classifiers are called the aggregated pairwise classifiers on a single tangent space, but with multiple tangent PC subspaces. Similarly, we first map all training shapes into a tangent space at the overall mean $\bar{q}$. Next, we compute all possible pairwise covariance matrices of classes $i$ and $j$, $i \neq j = 1, \ldots, K$. Then we can obtain $M = \binom{K}{2}$ sets of tPC coefficients, all of which dimensions are reduced to $r$. Once we have a test shape $x$, all pairwise log-likelihoods, $l_{\bar{q}}^{i,j}(x; \hat{\mu}_k, \hat{\Sigma}_k)$, are computed respectively and then aggregated by

$$\bar{l}_{\bar{q}}\left(x; \hat{\mu}_k, \hat{\Sigma}_k\right) = M^{-1}\sum_{i<j} l_{\bar{q}}^{i,j}\left(x; \hat{\mu}_k, \hat{\Sigma}_k\right). \tag{3.2}$$

We can also use $\hat{\Sigma}_P$ for LDA. The class with the largest aggregated log-likelihood is finally chosen as the classification result of the test shape $x$. In other words, we aggregate results from pairwise PC spaces.

To sum up, there are three choices in the outlined classification procedures using the Gaussian models: (1) LDA vs. QDA, (2) single vs. aggregated likelihoods from pairwise comparisons, and (3) the dimensionality of the PC space. Traditionally, many multivariate problems with respect to the first choice have been dealt with. For the second choice, the aggregated procedures are expected to lead to more robust classification performance, however it could be computationally more expensive. Determining optimal tangent PC dimension is not trivial, and requires an extensive search across various classification problems or datasets. In general, we aim to achieve a low-dimensional Euclidean representation of shape data via tPCs with decent performance of classification. Through the following empirical studies with algal shape data, these classification approaches are compared and some practical considerations are addressed.

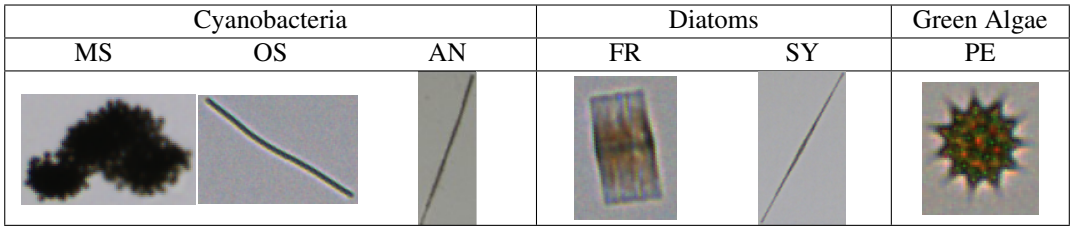


Figure 5: *A sample image for each species showing its shape.*

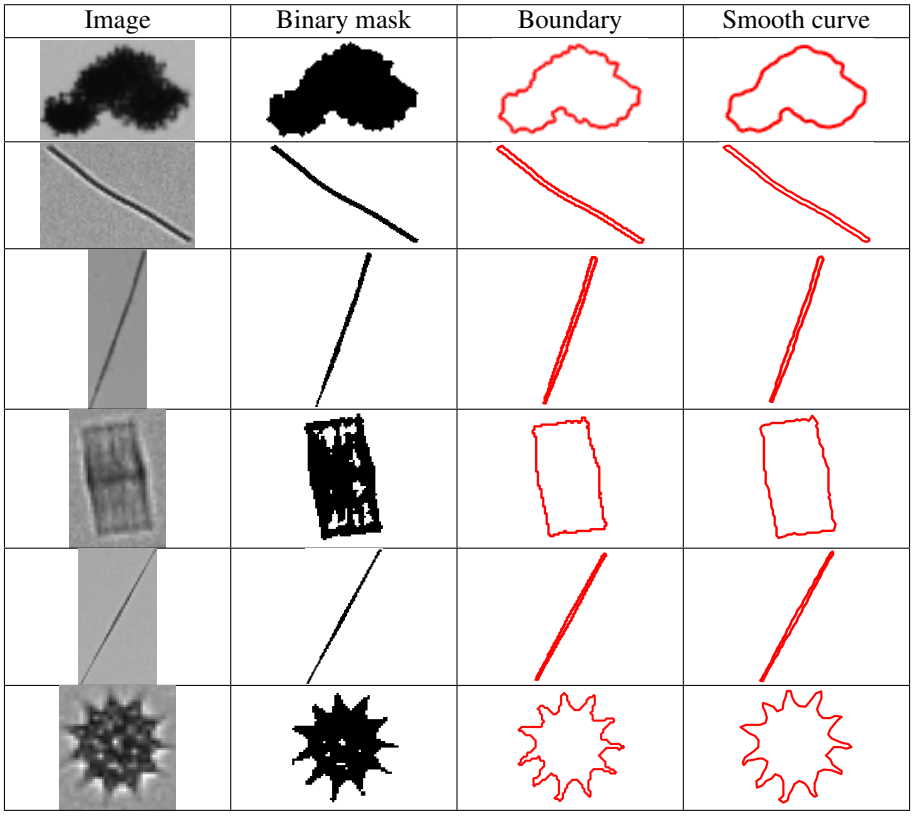


Figure 6: *An example of the preprocess to obtain a closed curve from the original image.*

## 4. Empirical studies

In this section, we provide the application of the classification methods previously described to shape data of algae in water. For the nearest classifiers, we focus on the overall performance and the benefit of using the elastic distance rather than the non-elastic one. For the models based on multivariate Gaussian distributions, we compare the overall accuracy as well as investigate the trend of the performance with various choices of the tPC dimension.

Table 1: Average classification rates (%) over 20 random splits of three distance-based methods for algal shape

| | Nearest neighbor | | | | | | | | Nearest |
|---|---|---|---|---|---|---|---|---|---|
| | $k = 1$ | $k = 3$ | $k = 5$ | $k = 7$ | $k = 9$ | $k = 11$ | $k = 13$ | $k = 15$ | mean |
| **Elastic** | 83.36 | 84.02 | **84.22** | 84.03 | 83.63 | 83.57 | 83.28 | 82.91 | 73.91 |
| | (1.14) | (1.08) | (1.08) | (0.88) | (0.90) | (0.86) | (0.86) | (0.87) | (1.10) |
| **Non-elastic** | 76.92 | 76.33 | 75.72 | 75.00 | 74.24 | 73.68 | 72.80 | 72.18 | |
| | (0.88) | (0.82) | (1.09) | (0.96) | (1.08) | (1.10) | (1.06) | (1.13) | |

## 4.1. Data description

The dataset we applied for in this work was first obtained from images, which were collected using a FlowCAM in a project by the Korea Water Resources Corporation (K-water) in 2015. As described in the motivating subsection, the images were captured and collected in the major rivers in South Korea, and were used for classification by Park *et al.* (2019). This dataset consists of algal images from six different species: Microcystis (MS), Oscillatoria (OS), Anabaena (AN), Fragilaria (FR), Synedra (SY), and Pediastrum (PE). There are 450 different sample images for each species except for AN, which has 321 images. Figure 5 shows one sample image for each of the six algal species. Microcystis (MS), Fragilaria (FR), and Pediastrum (PE) seem to have unique shapes, so they are relatively easy to visually discriminate. On the other hand, Oscillatoria (OS), Anabaena (AN), and Synedra (SY) all have similar long and thin shapes, and thus, the presented scientific procedures are necessary to more accurately distinguish these species via its shape.

## 4.2. Data preprocessing

In the preprocess to extract a curve from an image of an object of interest, the first step was to segment its region by converting the original image to one with a binary mask. After choosing an appropriate threshold pixel value for each image, we replaced all pixel values to either one or zero. Then, we extracted all coordinate values $(x_i, y_i)$, $i = 1, \ldots, m$ of the black mask outline as the third picture in Figure 6, where $m$ is the desired number of points to represent its shape. We interpolated the boundary points using a continuous function, such as splines, and included the smoothing step, if necessary, to avoid too wiggly a boundary or to reduce the effect of noise.

## 4.3. Classification results of algal shape

Here, we present classification results of various approaches using the elastic shape from algal images. The dataset consists of two-dimensional closed curves of the $(x, y)$ coordinates for 100 points, which trace the outline of the algae to represent each shape. The data were randomly split into 60% as a training set (270 samples for each species and 193 for AN) and the other 40% as a test set (the remaining 180 samples and 128 for AN). For evaluation and comparison among various classification methods, the (test) average classification rate over 20 random splits were used. In this subsection, we considered the multiclass classification problem for shapes of 6 algal species and compared the results separately: (1) among distance-based classification methods in Table 1 and (2) model-based approaches in Figure 7.

When comparing the three distance-based classification procedures using the cross-validation with 20 random splits: $k$-nearest neighbor (both elastic & non-elastic) and the nearest mean classifiers, which is the nearest neighbor classifiers based on the elastic distance, had overall higher average accuracy, more than 80%, which was significantly better than the nearest neighbor methods based on the non-elastic distance and the nearest mean classifier. Among the different choices of the number

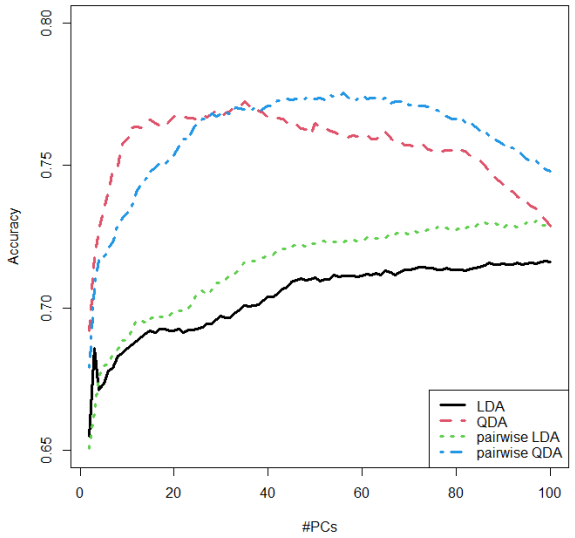


Figure 7: *Average classification rates (%) over the same 20 random splits of four model-based methods when using different number of tangent principal components for algal shape.*

of neighbors $k$ for the elastic distance, the average accuracy increased when we increased $k$ from one. The $k = 5$ nearest neighbor classifier produced the most accurate classification result of 84.22%, and its performance started to deteriorate when more neighbors were used. Similar patterns of the average accuracy for the different choices of $k$ in the nearest neighbor methods based on the non-elastic distance were shown with the highest accuracy of about 77%. The nearest mean classifier showed the classification result of around 74%. This suggested that a local classifier akin to a few nearest neighbor classifiers, was better suited for this dataset, and that the elastic distance had strength compared to the non-elastic distance for shape classification.

When comparing the four model-based classification procedures using the cross-validation with the identical 20 splits for the distance-based methods: LDA, QDA, pairwise LDA, and pairwise QDA; the highest average accuracy of QDA and pairwise QDA were 77.25% (with a standard deviation of 1.22%), and 77.55%(1.11%) when using the first 35, 56 tPCs, respectively as shown in Figure 7. As the number of tPCs increased, the average accuracy of LDA and pairwise LDA consistently increased while the accuracy of QDA and pairwise QDA increased up to a certain number of PCs, and then started to decrease. Overall, the QDA classification methods showed better performance than the LDA-type approaches. The discrepancy of their performance was large when a smaller number of tPCs were used. Moreover, the pairwise methods, by aggregating multiple likelihoods from pairwise PC spaces tended to have slightly more robust classification results for various choices of tPCs.

The overall classification accuracy of the model-based approaches was worse than that of the nearest neighbor classifiers based on the elastic distance. This might be caused by some characteristics that the probability models have: (1) classification by the global structure of the estimated distributions on the reduced dimensional space and (2) some distortion between an approximated tangent space and the original shape space.

## 5. Discussion and conclusions

We introduced several classification approaches for shape data based on the elastic shape analysis framework, and applied them to algal identification via their shape. Since we assumed the representation for algal shape as a continuous planar curve, we needed to overcome some challenges for analysis: (1) invariance properties of shape, (2) nonlinearity of its representation space, and (3) high dimensionality. Based on the elastic metric with the square-root velocity function, we could define the shape distance that led to better classification results. Based on the linearization of the data in tangent spaces and the dimension reduction by PCA on the tangent space, we could build probability models on the lower-dimensional Euclidean space to be used for shape classification.

However, there are some practical issues for the presented classification approaches for elastic shape. First, the $k$-nearest neighbors classifier requires computation of pairwise geodesic distances between all training and test cases; and each computation involves a complex nonlinear registration problem. As a result, this approach is computationally inefficient when there are many training cases in the data. When the number of nearest neighbors $k$ is greater than one, class ties can exist among the k neighbors. Although one can avoid ties using an odd number k in binary classification, there is a need for additional tie breaking rules in the multiclass classification problem. When this situation happened to us, we broke the tie by reducing the neighborhood size stepwise from $k$ to $k-1,\ldots,1$ (if necessary) until there were no ties (Weinberger and Saul, 2009).

In addition, there is room for improvement of the model-based classification. A single projection at the overall mean shape might result in distorted likelihoods for classification. If we develop more local linearization or better intrinsic ways to build statistical models, we can enhance the classification performance for shape. For future work, we will seek alternatives to linearization and dimension reduction procedures that can improve classification performance. Further enhancements, beyond the shape of algae, can be developed in methodology to incorporate other features such as texture (pixel or voxel values), inside the object of interest. We will consider statistical models applicable to various types of data that represent morphological features of an object in images or videos.

## Acknowledgement

This study was supported by K-water (Korea Water Resources Corportion). This work was supported by a National Research Foundation of Korea (NRF) grant funded by the Korean government (RS-2022-00167077) and a INHA UNIVERSITY Research Grant.

## References

Besl PJ and McKay ND (1992). Method for registration of 3-D shapes, *Sensor Fusion IV: Control Paradigms and Data Structures*, **1611**, 586–607.

Cho MH, Kurtek S, and MacEachern SN (2021). Aggregated pairwise classification of elastic planar shapes, *The Annals of Applied Statistics*, **15**, 619–637.

Coltelli P, Barsanti L, Evangelista V, Frassanito AM, and Gualtieri P (2014). Water monitoring: Automated and real time identification and classification of algae using digital microscopy, *Environmental Science: Processes & Impacts*, **16**, 2656–2665.

Dauta A, Devaux J, Piquemal F, and Boumnich L (1990). Growth rate of four freshwater algae in relation to light and temperature, *Hydrobiologia*, **207**, 221–226.

Dryden IL and Mardia KV (2016). *Statistical Shape Analysis: With Applications in R*, Wiley.

Grenander U and Miller MI (1998). Computational anatomy: An emerging discipline, *Quarterly of*

*Applied Mathematics*, **LVI**, 617–694.

Joshi SH, Klassen E, Srivastava A, and Jermyn IH (2007). A novel representation for Riemannian analysis of elastic curves in $\mathbb{R}^n$. In *Proceedings of 2007 IEEE Conference on Computer Vision and Pattern Recognition*, Minneapolis, MN, 1–7.

Kendall DG (1984). Shape manifolds, procrustean metrics, and complex projective spaces, *Bulletin of the London Mathematical Society*, **16**, 81–121.

Kurtek S, Srivastava A, Klassen E, and Ding Z (2012). Statistical modeling of curves using shapes and related features, *Journal of the American Statistical Association*, **107**, 1152–1165.

Kurtek S, Su J, Grimm C, Vaughan M, Sowell R, and A Srivastava (2013). Statistical analysis of manual segmentations of structures in medical images, *Computer Vision and Image Understanding*, **117**, 1036–1050.

Malladi R, Sethian JA, and Vemuri BC (1996). A fast level set based algorithm for topology-independe nt shape modeling, *Journal of Mathematical Imaging and Vision*, **6**, 269–290.

Medina E, Petraglia MR, Gomes JGR, and Petraglia A (2017). Comparison of cnn and mlp classifiers for algae detection in underwater pipelines. In *Proceedings of 2017 Seventh International Conference on Image Processing Theory, Tools and Applications (IPTA)*, Montreal, QC, 1–6.

Michor PW and Mumford D (2006). Riemannian geometries on spaces of plane curves, *Journal of European Mathematical Society*, **8**, 1–48.

Mio W, Srivastava A, and Joshi SH (2007). On shape of plane elastic curves, *International Journal of Computer Vision*, **73**, 307–324.

Paerl HW and Otten TG (2013). Harmful cyanobacterial blooms: Causes, consequences, and controls, *Microbial Ecology*, **65**, 995–1010.

Pal S, Woods RP, Panjiyar S, Sowell ER, Narr KL, and Joshi SH (2017). A Riemannian framework for linear and quadratic discriminant analysis on the tangent space of shapes, In *Proceedings of Workshop on Differential Geometry in Computer Vision and Machine Learning*, Honolulu, HI, 726–734.

Park J, Lee H, Park CY, Hasan S, Heo T-Y, and Lee WH (2019). Algal morphological identification in watersheds for drinking water supply using neural architecture search for convolutional neural network, *Water*, **11**, 1338, 1–19.

Pizer SM, Jung S, Goswami D *et al.* (2013). Nested sphere statistics of skeletal models, In *Innovations for Shape Analysis: Models and Algorithms*, 93–115.

Robinson DT (2012). Functional data analysis and partial shape matching in the square root velocity framework (Ph.D thesis), Florida State University, Tallahassee, FL.

Srivastava A, Jermyn IH, and Joshi S (2007). Riemannian analysis of probability density functions with applications in vision, In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, Minneapolis, MN, 1–8.

Srivastava A, Klassen E, Joshi SH, and Jermyn IH (2011). Shape analysis of elastic curves in Euclidean spaces, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **33**, 1415–1428.

Srivastava A and Klassen EP (2016). *Functional and Shape Data Analysis*, Springer, New York.

Weinberger KQ and Saul LK (2009). Distance metric learning for large margin nearest neighbor classification, *Journal of Machine Learning Research*, **10**, 207–244.

*Received August 29, 2023; Revised October 29, 2023; Accepted December 24, 2023*