Characterizing Cell Shape Distributions Using k-Mode Kernel Mixtures

Ximu Deng, Anuj Srivastava
Department of Statistics,
Florida State University
Tallahassee, FL, USA
Email: xd15@fsu.edu, anuj@stat.fsu.edu

Rituparna Sarkar, Elisabeth Labruyère and Jean-Christophe Olivo-Marin
Bioimage Analysis Unit
Department of Cell Biology and Infection
Institut Pasteur, Paris, France
Email: {rituparna.sarkar, elisabeth.labruyere, jcolivo}@pasteur.fr

***Abstract*—This paper addresses the problem of characterizing statistical distributions of cellular shape populations using shape samples from microscopy image data. This problem is challenging because of the nonlinearity and high-dimensionality of shape manifolds. The paper develops an efficient, nonparametric approach using ideas from k-modal mixtures and kernel estimators. It uses *elastic shape analysis* of cell boundaries to estimate statistical modes and clusters given shapes around those modes. (Notably, it uses a combination of modal distributions and ANOVA to determine k automatically.) A population is then characterized as k-modal mixture relative to this estimated clustering and a chosen kernel (e.g., a Gaussian or a flat kernel). One can compare and analyze populations using the Fisher-Rao metric between their estimated distributions. We demonstrate this approach for classifying shapes associated with migrations of *entamoeba histolytica* under different experimental conditions. This framework remarkably captures salient shape patterns and separates shape data for different experimental settings, even when it is difficult to discern class differences visually.**

I. Introduction

Cellular morphogenesis during migration is an interesting topic of research in cell biology. Cell migration is a complex process influenced by changes in membrane dynamics, polarization and restructuring of the cytoskeleton. The highly deformable cell membrane aids the cells to adapt to a micro-environment, switch between motility patterns and induce positional advancement. This dynamic morphology introduces significant variation in cellular shapes adopted during its course of motion leading to various morphological modes of cell migration. For cell morpho-dynamics study, shape analysis has been primarily used for clustering [1], [2] and classification of migration patterns under different experimental condition [3], [4], [5], [6], [7], [8]. Both static shapes and shape evolution dynamics have been used in this problem area.

From a biological perspective, the focus here is on unicellular amoebas and some cancer cells. These organisms exhibit amoeboid migration patterns invading their surroundings or escaping from the originating tissues. Such amoeboid motions are characterized by alternate protrusions and retractions of the cell membranes, resulting from intra-cellular bio-physical changes and adhesion to the extracellular substrate. We focus on single-cell *Entamoeba histolytica* as the prototype organism to study cell motility. In the video frames, we will consider amoeba as 2D objects represented by their boundaries (simple, closed, planar curves). We employed an existing approach [9] designed explicitly for segmenting amoeboid cells from brightfield microscopy and extracting cell boundaries. Examples of extracted shape sequences from different experimental conditions are shown in Fig. 1.

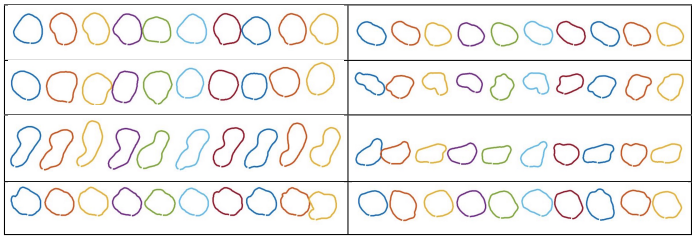


Fig. 1: Examples of extracted shape sequences of migrating amoeboid cells from four different settings. Each row exhibits shows two sets of shapes from an experimental class. From top to bottom, the classes are Fibronectin w/o inhibitor, Fibronectin with inhibitor, Glass w/o inhibitor, and Glass with inhibitor.

This discussion motivates the study of shapes as entire collections rather than individual shapes associated with different migration conditions. The underlying biological hypothesis is that different settings often result in statistically different distributions of shapes, even though some individual shapes may appear similar across settings. Consequently, we are interested in characterizing statistical distributions of cellular shapes from given samples. These shape distributions can, in turn, be used for inferences about the migration patterns under different micro-environments. Therefore, we aim to **characterize and compare statistical distributions of shapes** using shape samples obtained from microscopy data.

The problem of characterizing shape distributions can be approached in several ways. Let $\mathcal{S}$ represent a shape space, endowed with a metric d_s for quantifying shape differences. For example, $\mathcal{S}$ can be the Kendall's landmark-based shape space [10], elastic shape space of curves [11], or some other shape representation. In most approaches, $\mathcal{S}$ is a nonlinear and high-dimensional Riemannian manifold or a quotient space of such a manifold. Thus, standard (Euclidean) multivariate statistical models do not apply.

- **A Parametric Model**: A simple and popular idea is to define a parametric family, such as a Gaussian or its analog

adapted to the shape space $\mathcal{S}$, for capturing given shape variability. Given samples from an underlying distribution, one can estimate shape summaries [12], [13] (mean, tangent space covariance, etc.), perform tangent space PCA, and use the results to estimate distribution parameters. This approach may not be helpful if the underlying population is complex and diverse, *e.g.*, it is multimodal. The dynamic membrane of motile cells generates considerable variation in cellular shapes, which cannot be captured using a single cluster or class mean solely.

- **A Mixture Model**: One can treat a multimodal density as a mixture of unimodal distributions (each belonging to a simple parametric family) and use an EM-type algorithm to estimate mixture parameters [14]. One can also represent shapes as linear models [15] and estimate model parameters to characterize shapes in large datasets. The computational tools needed to analyze mixture models include the k-mean clustering, hierarchical clustering, and EM-type estimation algorithms [16], [17], [18], [19], [2]. There are several limitations to these mean-based approaches to shaping spaces. Firstly, computing the mean of a set of shapes is computationally expensive, especially when there is an additional outer loop for updating means, as in the k-mean or EM-type algorithm. Secondly, shape means or even shape medians denote an averaging result that smooths out features of the observed cellular shapes. As we demonstrate later, these shape means are not good representatives of the sampled shapes. Lastly, determining the optimal number of clusters or components in mixture models to capture the data variability is a challenging task by itself, and naívely selecting these parameters can significantly alter the subsequent analysis.
- **A Kernel method**: Nonparametric approaches have gained prominence in recent years because they don't require any assumption about the structure of the distribution. In fact, one can estimate the underlying *pdf* in a nonparametric form using a kernel $\mathcal{K}$ according to:

$$\hat{f}(s) = \frac{1}{n\epsilon} \sum_{i=1}^{n} \mathcal{K}_\epsilon(d_s(s, s_i)), \tag{1}$$

where $s, \{s_i\}$ denote a general shape and the sampled shapes, respectively, and ϵ is the kernel bandwidth. (Technically, we are assuming that $\mathcal{S}$ is finite-dimensional and $\mathcal{K}$ is a positive-definite kernel on $\mathcal{S}$.) While this approach is generic and robust, its computational cost is prohibitive in practice. In order to compare two such shape distributions, one needs to evaluate the shape metric d_s between every reference point s to every data point s_i, which can be overwhelming.

A. Our Approach – k-Modal Mixtures of Shapes

In this paper, we develop a novel approach that combines the strengths of mixture models and nonparametric estimation. We seek an efficient, kernel-based estimation where the **kernels are placed only at a few salient shapes**. The question is: What should be the locations in $\mathcal{S}$ for centering these kernels? The solution comes from mixture models that center distributions around a few pivot points. However, in contrast to mean or median-based mixture models, we use the **statistical modes** [20], [21], [22], [23]. These modes are statistically significant local maxima of the underlying *pdf*. The advantage of using modes (over means or medians) are numerous, including the availability of a very efficient technique for estimating them from the observed shapes, see Deng et al. [24]. We then define a kernel-based density estimator for only these modal shapes, according to:

$$\hat{f}(s) = \frac{1}{m\epsilon} \sum_{j=1}^{m} \mathcal{K}_\epsilon(d_s(s, s_j)), \tag{2}$$

where $\{s_j, j = 1, 2, \ldots, m\}$ is the set of modal shapes. Since $m << n$, we get efficiency and compactness in representing the shape population by mixtures. We term this expression as a **k-modal kernel mixture** estimate of the underlying shape density. We demonstrate a special case where the kernel is simply an indicator function so that the estimated distributions reduce to normalized histograms of the shapes relative to a fixed shape clustering. Lastly, we can analyze and compare estimated shape distributions associated with different shape classes using the Fisher-Rao metric [25].

To demonstrate the strengths of this framework, we apply this approach to cell populations associated with different experimental conditions.

The main contributions of paper are as follows:

1) It develops a novel, nonparametric representation for *pdf*s of shapes as a k-modal kernel mixture and derives an efficient procedure for estimating the *pdf*. The proposed estimate (Eqn. 2) is much more efficient to analyze and compare than the original nonparametric estimate (Eqn. 1) or a mixture of Gaussians. To the best of our knowledge, this is the first use of k-modal mixture distributions to represent and compare shape populations.
2) It uses a current framework for estimating shape modes from sample data to develop a k-modal mixture estimate. Here, one estimates modes of *pdf* directly from the data, without first resorting to the estimation of *pdf*. An important aspect is that the number of components k and the bandwidth ϵ are determined automatically from the data using ANOVA.
3) It further simplifies *pdf*s as histograms of cluster memberships when using an indicator function as a kernel. This results in a fast comparison of distributions across shape classes. The paper uses the Fisher-Rao metric to compare estimated shape distributions, but one can use other metrics instead.
4) The paper applies these techniques to comparisons of migration patterns of *Entamoeba Histolytica* under different migration conditions, resulting in high classification performance. This is the first result in the field that successfully distinguishes (with rates $\sim 97\%$) biological classes using comparisons of statistical shape distributions.

II. Shape Distributions as Modal Mixtures

In this section, we specify the shape space of planar curves and use its geometry to define relevant statistics – modes, means, etc. – and characterize *pdf*s on this shape space. The key idea is to compute **shape modes** from the sample shapes and use them to describe the underlying shape distribution as a k-modal kernel mixture.

A. Background: Elastic Shape Analysis

We define a shape space $\mathcal{S}$ and consider a *pdf* f on $\mathcal{S}$ associated with a shape population. Given a set of closed planar curves $\beta_1, \beta_2, \ldots, \beta_n$, each representing an observed cell boundary, we treat their shapes as samples from f on $\mathcal{S}$. Our goal is then to estimate the modes of f from this sample data and use these modes to characterize dominant shapes in the data. We start with a brief introduction of elastic shape analysis [11], [26], [27] used for comparing cellular shapes. This elastic shape analysis has been used extensively in computer vision [28], biology [29], bioinformatics [30], and functional data analysis [31].

In this approach, a planar closed curve $\beta : \mathbb{S}^1 \to \mathbb{R}^2$ is represented by its Square-Root Velocity Function (SRVF) $q : \mathbb{S}^1 \to \mathbb{R}^2$ given by: $q(t) = \frac{\dot{\beta}(t)}{\sqrt{|\dot{\beta}(t)|}}$. The use of SRVF greatly simplifies shape analysis of curves, especially in invariance to kinematics (rotation, translation, scaling, and re-parameterization of β). Let $[q]$ the set of all possible rotations and re-parameterizations of SRVF q after rescaling q to have the unit $\mathbb{L}^2$ norm. The set of all shapes is denoted by $\mathcal{S} = \{[q] | q \in \mathbb{S}_\infty\}$. $\mathcal{S}$ is an infinite-dimensional, nonlinear space that limits our ability to perform traditional statistical analysis. Researchers have developed tools to study shapes as elements of $\mathcal{S}$. For any two shapes, one can compute a geodesic path between them and use the geodesic length as the shape distance d_s. Given a set of shapes, one can compute their mean and perform tangent PCA analysis for dimension reduction. For shapes $\{[q]_1, [q]_2, \ldots, [q]_n\}$, their Karcher or Fréchet mean is defined by the quantity: $[\hat{\mu}] = \text{argmin}_{[q] \in \mathcal{S}} \sum_{i=1}^n d_s([q], [q]_i)^2$. This mean is typically estimated using a gradient-based algorithm that requires computing a *shooting vector* from the current $[\hat{\mu}]$ to each $[q_i]$ in every iteration. Thus, this calculation is computationally expensive, especially in the elastic shape analysis, which uses the Dynamic Programming algorithm for registering the points across curves. We note that the resulting statistical summaries – means, shape distances, PCA, etc. – are invariant to global scale, rigid motions, and parameterizations, and are instrumental in separating cell kinematics from its morphology.

Before we proceed further, we point out that $\mathcal{S}$ is an infinite-dimensional space, and one cannot simply integrate a positive function on $\mathcal{S}$ or an open subset of $\mathcal{S}$ in a classical way. One approach to handle this issue is to model curves as realizations of a stochastic process. We bypass this problematic issue by considering a non-standard interpretation of *pdf*s. We will not insist on the *pdf* having an integral of one on $\mathcal{S}$. Instead, we will choose a different yet consistent way of normalizing these "*pdf*s". This normalization will be consistent across datasets and *pdf*s to facilitate a comparison of shape distributions across shape classes and experiments.

B. Nonparametric k-Modal Density Estimator

Given n closed curves $\{\beta_i,\ i = 1, ..., n\}$, we treat their shapes $\{[q_i] \in \mathcal{S}\}$ as samples from an underlying density f on $\mathcal{S}$. As discussed earlier, one can choose a parametric or a nonparametric form of f for statistical analysis.

The classic nonparametric approach, using a kernel, uses an expression of the type given in Eqn. 1. One needs to specify a kernel and a "Gaussian" kernel is a popular choice. It takes the form:

$$\mathcal{K}_\epsilon([q], [q_i]) = \exp(-d_s([q], [q_i])^2/\epsilon) \ .$$

However, the use of this estimator is cumbersome in practice. In order to evaluate $\hat{f}$ for any $[q]$, we need to compute the shape metrics (d_s) n times, where n is the sample size. In case n is large, and one needs to evaluate $\hat{f}$ for many shapes, this computation becomes a bottleneck. Another approach is to represent $\hat{f}$ as a mixture of distributions from a parametric family (e.g., Gaussian). Using a mixture of Gaussians (with Gaussian defined appropriately for the shape space $\mathcal{S}$) is also problematic, at least computationally. As noted earlier, the computation of a mean of shapes on $\mathcal{S}$ is expensive, making either k-mean clustering or an EM-type algorithm ineffective on a large dataset.

Our approach combines the strengths of mixture models and nonparametric approaches by developing a k-modal mixture kernel distribution. We express the desired *pdf* as a mixture of k-kernel functions, each centered around a significant mode of the underlying density (see Eqn. 2). In order to estimate and analyze this density, we first estimate statistical modes $\{[q_j]\}$ from sample shape data. Then, we construct $\hat{f}$ around those modes. (Note that this approach is in contrast to [22] where they estimate $\hat{f}$ nonparametrically first, then seek the modes of the estimate distribution using a gradient-based optimization of $\hat{f}$.)

We will investigate a particular case where the kernel is simply an indicator function:

$$\mathcal{K}_\epsilon(d_s(s_i, s_j)) = I(d_s(s, s_j) < \epsilon) = I(s \in \mathcal{B}_\epsilon(s_j)),$$

where $\mathcal{B}_\epsilon(s_j)$ is a ball of radius ϵ centered at $s_j \in \mathcal{S}$. With this choice of kernel, $\hat{f}$ is completely specified by a histogram (frequency count) of a shape sample in each cluster. This simplifies the original formulation, albeit at some loss of information, and provides an efficient way to compare shape distributions.

C. k-Mode Clustering of Shape Data

This section describes a recent discrete, nonparametric approach for estimating shape modes and clusterin shapes [24]. It is an iterative approach, similar to k-means clustering, where one makes an initial guess about k modes and iteratively refines the clustering. Using an elastic shape metric d_s, it defines ϵ-neighborhoods in the shape space $\mathcal{S}$ and shortlist

Algorithm 1 Shape Mode Estimation and Clustering

Require: Closed curves β_i, $i = 1, ..., n$. Compute their shape representations $[q_i] \in \mathcal{S}$, $i = 1, 2, \ldots$

1: For each shape $[q_i]$, find it's neighbors:

$$\mathcal{N}_i = \{[q_j] : d_s([q_i], [q_j]) < \epsilon\}, i \neq j \quad (3)$$

Let $|\mathcal{N}_i|$ denote the number of neighbors of $[q_i]$.

2: Find the k^{th} mode $[q_{M_k}]$ as follows: Select the set $A = \{[q_j] || \quad |\mathcal{N}_j| = \max_i(|\mathcal{N}_i|)\}$ and set $[q_{M_k}] = \min_{[q_j]\in A} \left(\sum_{[q_i]\in\mathcal{N}_j} d_s([q_j], [q_i])\right)$.

3: if $|\mathcal{N}_{M_k}| < 2$, we label $[q_{M_k}]$ an outlier, else it is called mode. Remove $[q_{M_k}]$ and its neighbors $\mathcal{N}_{M_k}$ from the data set.

4: Repeat Step 1 to Step 3 until each curves is defined either as a mode or a neighbor or an outlier.

shapes that are central and have the most neighbors. A critical issue – How to automatically select the threshold ϵ? – is resolved using a combination of ANOVA and empirical mode distribution. In each iteration, the given shapes are assigned to the nearest shape mode, automatically clustering the data. The shapes that are far away from all modes are labeled as outliers. Algorithm 1 summarizes the main steps of this procedure. This discrete and nonparametric approach is an efficient solution to seeking sample modes. We will present several illustrations of the algorithm later. Once we have sample modes and data is clustered around these modes, we can express the estimated shape *pdf* using Eqn. 2.

D. Fisher-Rao Metric of Comparing Shape Distributions

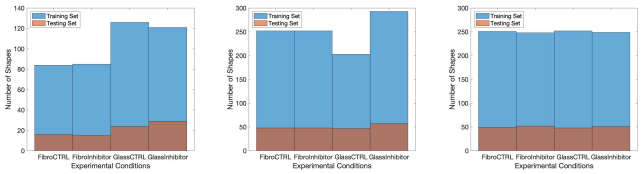
Once we have estimated *pdf*s for different shape populations, we can use them to compare these populations. We will use the Fisher-Rao metric (FRM) for this purpose. There are several quantities for comparing densities, such as the Kullback-Leibler divergence, but often they don't form proper distances. We will compute the FRM using the square-root transformations as follows. Let $g = \sqrt{\hat{f}}$ denote the point-by-point, positive square root of a probability density function f. Since $\hat{f}$ is a *pdf* and integrates to a constant, the $\mathbb{L}^2$ norm of g is also a constant. Hence, g is an element of the positive orthant $\mathbb{S}^+_\infty$ of the infinite-dimensional sphere $\mathbb{S}_\infty$, and referred to as the *half density* of $\hat{f}$. It is well known that the FRM for probability densities transforms to the $\mathbb{L}^2$ metric under the square-root mapping, up to a constant [25]. Thus, given any two density functions $\hat{f}_1$, $\hat{f}_2$, the FRM between them is:

$$d_f(\hat{f}_1, \hat{f}_2) = cos^{-1}\left(\int_{\mathcal{S}} \sqrt{\hat{f}_1([q])}\sqrt{\hat{f}_2([q])}d[q]\right). \quad (4)$$

In case the densities are expressed as histograms with respect to fixed shape clusters, the FRM is simply cosine inverse of the inner-product of the weighted histogram vectors.

III. Experimental Validation

This section presents some experimental results from our method applied to several real cell-shape datasets. Cell shapes in different biological experimental protocols are governed by underlying conditions that dictate specific behavior. Deng et al. [24] focused on individual dynamical shapes using the dynamic shape and kinematic features. However, it is possible that when some individual shapes are similar under different settings, this method may fail and result in degraded performance. In contrast, our k-modal kernel mixtures (Eqn. 2) characterize the complete statistical distributions of shapes rather than comparing individual shapes.



(a) Experiment 1 (b) Experiment 2 (c) Experiment 3

Fig. 2: The three histograms display the numbers of cell shapes for different datasets. Each histogram represents number of shapes per class for four biological conditions used in the experiments. Blue bars refer to training set and red refer to testing set.

To evaluate classification performance using the proposed framework, we perform experiments on brightfield microscopy images of *E. histolytica*. The data were obtained for four different biological conditions, where amoebas are seeded on glass and fibronectin with or without a rho-kinase inhibitor. These four biological classes are denoted here as:'FibroCTRL', 'FibroInhibitor', 'GlassCTRL', and 'GlassInhibitor'. The shape sequences are extracted from the microscopy images using *MultiCell-Net* [9]. We perform three sets of experiments with this data, and for each experiment, we randomly select shapes from different cell sequences. As shown in Fig. 2, the sizes of the four classes are unbalanced for both the training set and testing, and the distributions of shapes are also different in these three experiments.

A. Clustering and Shape Distribution

Experiment 1: Here we select 500 cell shapes from 10 different cell migration sequences, *i.e.*, 50 shapes from each sequence, with sequences taken from four classes. Fig. 3 shows some examples at the top. We divide them randomly into a training set (417 shapes) and testing set (83 shapes). The distribution of shapes per class are shown in Fig. 2(a). For the training set, we study the influence of ϵ on the number of modes and the *F-statistic*. Fig. 3(a) shows the changes in all-modes, *significant* modes, and the outliers as ϵ changes. Fig. 3(b) displays the relation between ϵ and the number of significant modes and the *F-statistic*. In Fig. 3(b), the peak of the blue curve is at $\epsilon_M = 0.2218$ and that of the orange curve is at $\epsilon_F = 0.2776$. The optimal $\epsilon = 0.5\epsilon_M + 0.5\epsilon_F = 0.2497$, which yields **five** distinct clusters. An MDS plot (Fig. 3(d)) shows shapes as points in a 2D plane with colors showing the groupings, and the five clusters are well separated from each other. The Fig. 3(c) shows the overall Karcher mean

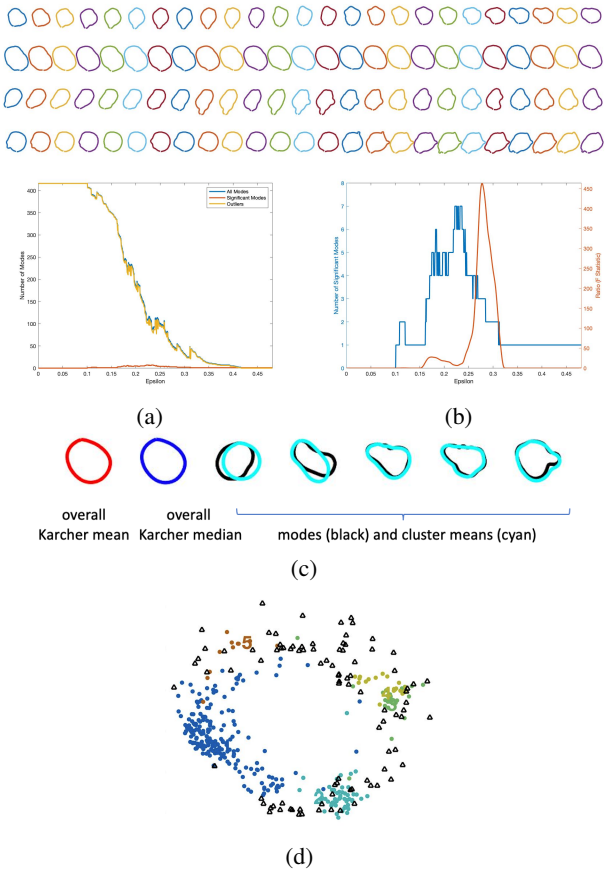


Fig. 3: Top: Sample shapes used in the Experiment 1. Plot (a): shows number of modes vs. ϵ. The blue curve refers to total number of modes, yellow curve shows the outliers and red indicates *significant* modes. Plot (b): the blue curve shows number of *significant* modes and orange curve indicates *F-statistic* w.r.t ϵ. Plot (c): Overall Karcher mean (red), overall Karcher median (blue), modes (black) and cluster means (cyan). Plot (d): MDS plot.

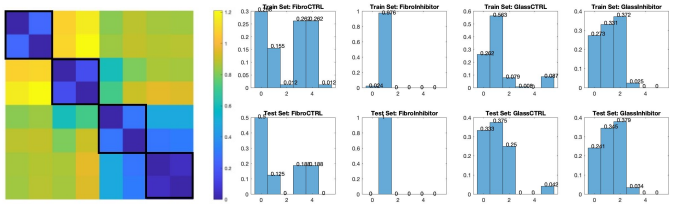


Fig. 4: Subfigure (a): It shows the Fisher metric for eight histograms (including both training and testing datasets) which are shown in the bottom plot. Subfigure (b): The first row shows the histograms obtained using the clustering results for four different experimental conditions for the training set. The second row shows the histogram for the testing set using the cluster modes obtained for the experiment.

(red), overall Karcher median (blue) and the estimated modes (black) overlaid on the group means (cyan). Due to significant variation within the clusters, the overall Karcher mean and median lose critical shape features while mode estimation retains the features relevant to each cluster.

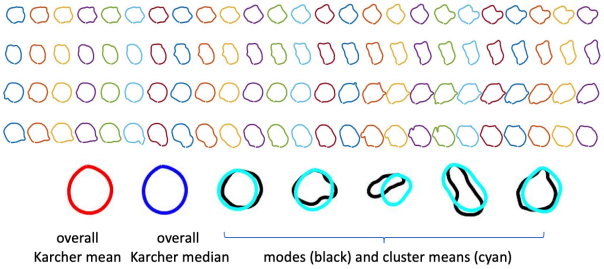


Fig. 5: Top: Sample shapes used in the Experiment 2. Bottom: Overall Karcher mean (red), overall Karcher median (blue), modes (black) and cluster means (cyan).

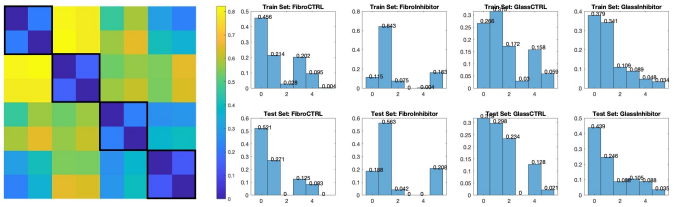
Fig. 4 shows the histograms of cluster memberships of different classes and compares them using the FRM. In Fig. 4(b), each plot displays the estimated *pdf* under the flat kernel as normalized histograms. The x-axis is the cluster labeling from zero to five, where the zeroth bin denotes the outliers. The first row shows the clustering results for the training set, and the second row shows the same for the testing set. Fig. 4(a) visualizes the FRM, an 8×8 matrix, for the eight shape distributions shown in the corresponding histogram plots. Every two rows/columns are the FRM for the training set and testing corresponding to the same experimental condition/category. According to this plot, the FRM between testing and training distribution in the same class is the smallest compared to the training set from other classes. That is, the probability densities within the classes are more similar than across the classes.

Experiment 2: In this experiment, the dataset contains 1200 cell shapes from 24 different cell migration sequences - 50 shapes from each sequence. Some example shapes are shown at the top of Fig. 5. We divide them randomly into a training set (1000 shapes) and a testing set (200 shapes). The distribution of shapes per class is given in Fig. 2(b). For this data, we obtain $\epsilon_M = 0.2218$, $\epsilon_F = 0.2776$ and the optimal $\epsilon = 0.2497$. For this ϵ, we discover **five modes**. Fig. 5 also shows the estimated modes and compares them with the cluster means. Considering the relatively larger variation in this dataset, the difference between the estimated modes and cluster means is larger here than in Experiment 1. In Fig. 6 (b), the estimated *pdfs* are found to be similar for within-class training and testing sets, except for the class 'GlassInhibitor,' where the largest group is cluster 0 for the training set while it is cluster 1 for the testing set. Noting that cluster 0 is the group for outliers, the two densities can be considered similar if the outliers are ignored. Fig. 6 (a) shows class 'FibroCTRL', 'FibroInhibitor' 'GlassInhibitor' have smaller FRM between the training and testing set.

Experiment 3: Similar to Experiment 2, we select 1200 cell shapes from 24 different cell migration sequences - 50 shapes in each sequence. These cell sequences are completely different from those used in Experiment 2. The numbers of

TABLE I: Classification performance

Method	**Shape Distribution**	2D Shape + Conv NN	TSRVF+PCA + Conv NN	TSRVF+PCA +VAR + Conv NN [6]	Kinematics + Conv NN	TSRVF+PCA+VAR +Kinematic+Conv NN[6]
Class. Rate (%)	**97.5**	41.5	34.1	83.5	76.9	84.9



(a) (b)

Fig. 6: Same as Fig. 4, but for Experiment 2.

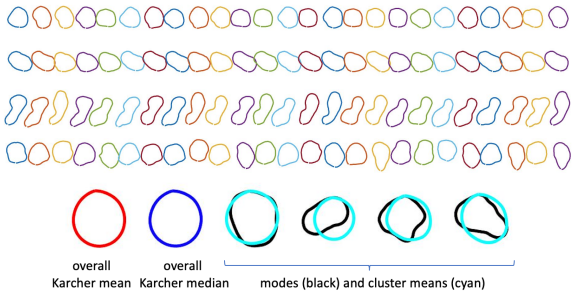
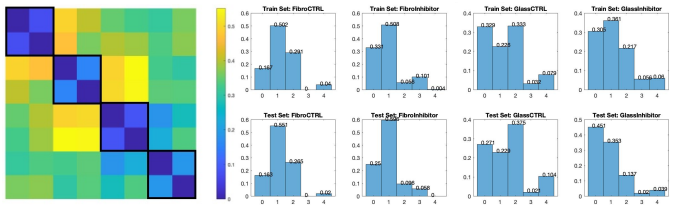


Fig. 7: Same as Fig. 5, but for Experiment 3.

shapes per class for training and testing are shown in Fig. 2. Here we find $\epsilon_M = 0.2119$, $\epsilon_F = 0.2432$ and the optimal being $\epsilon = 0.2276$ resulting in **four modes**. Fig. 7 also shows the estimated modes and their improvements over cluster mean shapes. Fig. 8(b) shows that estimated *pdf*s (normalized histograms) of the first three categories look similar between the training and testing set. Again, for the GlassInhibitor class, the histograms are slightly different, and the shape distribution per cluster is significantly different from the other classes. This is also evident from the FRM matrix in Fig. 8(a).



(a) (b)

Fig. 8: Same as Fig. 4, but for Experiment 3.

B. Classification Performance

Besides demonstrating clustering and labeling, we also quantify our shape-distribution-based classification performance. Table. I shows the classification performance obtained using the dataset used in Experiment 3. In every run, we randomly split the data into the training set (1000 shapes) and testing set (200 shapes). We use the classification results from the training set to find the nearest training distribution for each testing distribution by selecting the smallest value of FRM. The average accuracy is found to be 97.5% for 20 runs, demonstrating that our approach is highly effective. We compare our method with classification performance using 2D shape vectors and TSRVF-PCA shape features with SOTA classifiers. In this case, we use the shape coordinates (or TSRVF-PCA vectors) as input to a convolutional neural network (CNN) to classify individual shapes rather than as shape distributions. The architecture consists of three convolutional layers (kernel size 3) sequentially followed by Batch-Norm, ReLU activation, dropout, and max-pooling. This network is trained using cross-entropy loss for 90 epochs with batch size 32 and learning rate 0.01.

To distinguish migration patterns in different experimental conditions, we also list classification results using shape dynamics, cell kinematics, and a combination of the two. The results were computed using the features described in [6] but on a larger dataset and using the architecture mentioned above. Here we employ the distance vector computed over training and test data, unlike that in [6], where the distance vector is computed only from the training data. Nevertheless, the higher classification accuracy using shape distribution conveys that the shape variations are distinctly different in the four experimental conditions, and this is captured more efficiently using our k-mode clustering approach. The confusion matrix for classification is presented in Table II which demonstrates that our method performs well with much higher classification accuracy and fewer false positives.

TABLE II: Confusion matrix for multi-class classification problem.

	Our Approach				Using Individual Shapes				Using Shape dynamics [6]			
	GC	GI	FC	FI	GC	GI	FC	FI	GC	GI	FC	FI
GC	**0.9**	0.1	0	0	**0.75**	0.14	0.06	0.04	**0.88**	0.11	0	0
GI	0.2	**0.7**	0	0.1	0.43	**0.49**	0.00	008	0.02	**0.79**	0.16	0.02
FC	0 .1	0	**0.9**	0	0.12	0.49	**0.35**	0.04	0	0	**0.90**	0.10
FI	0	0	0	**1**	0.42	0.38	0.09	**0.09**	0.21	0	0.05	**0.73**

IV. Conclusion

This paper introduces an efficient approach for characterizing statistical distributions of cellular shapes in a population using kernel-based, k-modal mixtures. This approach results in a discrete, non-parametric representation of a shape distribution and a fast comparison of distributions across categories using the FRM. Because of its generality, this characterization can be used in broad biological, bioinformatics, computer vision, and other domains as a general tool in statistical shape analysis. In future work, one can investigate using other kernels, e.g., Gaussian-type kernels, on shape manifolds to obtain performance improvements.

REFERENCES

[1] I. Ahonen, V. Härmä, H.-P. Schukov, M. Nees, and J. Nevalainen, "Morphological clustering of cell cultures based on size, shape, and texture features," *Statistics in Biopharmaceutical Research*, vol. 8, no. 2, pp. 217–228, 2016.

[2] Y. T. Maeda, J. Inose, M. Y. Matsuo, S. Iwaya, and M. Sano, "Ordered patterns of cell shape and orientational correlation during spontaneous cell migration," *PloS one*, vol. 3, no. 11, p. e3734, 2008.

[3] A. C. Dufour, T.-Y. Liu, C. Ducroz, R. Tournemenne, B. Cummings, R. Thibeaux, N. Guillen, A. O. Hero, and J.-C. Olivo-Marin, "Signal processing challenges in quantitative 3-d cell morphology: More than meets the eye," *IEEE Signal Processing Magazine*, vol. 32, no. 1, pp. 30–40, 2014.

[4] D. Imoto, N. Saito, A. Nakajima, G. Honda, M. Ishida, T. Sugita, S. Ishihara *et al.*, "Comparative mapping of crawling-cell morphodynamics in deep learning-based feature space," *PLoS Computational Biology*, vol. 17, no. 8, p. e1009237, 2021.

[5] X. Deng, R. Sarkar, E. Labruyere, J.-C. Olivo-Martin, and A. Srivastava, "Modeling shape dynamics during cell motility in microscopy videos," in *International Conference on Image Processing, ICIP*, October 2020.

[6] X. Deng, R. Sarkar, E. Labruyere, J.-C. Olivo-Marin, and A. Srivastava, "Dynamic shape modeling to analyze modes ofmigration during cell motility," *arXiv preprint arXiv:2106.05617*, 2021.

[7] A. Medyukhina, M. Blickensdorf, Z. Cseresnyés, N. Ruef *et al.*, "Dynamic spherical harmonics approach for shape classification of migrating cells," *Scientific Reports*, vol. 10, no. 1, pp. 1–12, 2020.

[8] F. L. Kriegel, J. Köhler, R.and Bayat-Sarmadi, S. Bayerl, A. E. Hauser *et al.*, "Cell shape characterization and classification with discrete fourier transforms and self-organizing maps," *Cytometry Part A*, vol. 93, no. 3, pp. 323–333, 2018.

[9] R. Sarkar, S. Mukherjee, E. Labruyère, and J.-C. Olivo-Marin, "Learning to segment clustered amoeboid cells from brightfield microscopy via multi-task learning with adaptive weight selection," in *2020 25th International Conference on Pattern Recognition (ICPR)*, 2021, pp. 3845–3852.

[10] I. L. Dryden and K. Mardia, *Statistical Shape Analysis*. John Wiley & Son, 1998.

[11] A. Srivastava and E. Klassen, *Functional and Shape Data Analysis*. Springer Series in Statistics, 2016.

[12] H. Le, "Locating frechet means with application to shape spaces," *Advances in Applied Probability*, vol. 33, no. 2, pp. 324–338, 2001.

[13] A. Srivastava, S. H. Joshi, W. Mio, and X. Liu, "Statistical shape anlaysis: Clustering, learning and testing," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 27, no. 4, pp. 590–602, 2005.

[14] R. Hosseini and S. Sra, "Matrix manifold optimization for gaussian mixtures," in *Advances in Neural Information Processing Systems*, vol. 28, 2015.

[15] M. Mirshahi, V. Partovi-Nia, and M. Asgharian, "A model selection approach to hierarchical shape clustering with an application to cell shapes," *bioRxiv*, 2020.

[16] L. Tweedy, B. Meier, J. Stephan, D. Heinrich, and R. G. Endres, "Distinct cell shapes determine accurate chemotaxis," *Scientific Reports*, vol. 3, p. 2606, 2013.

[17] Z. Yin, H. Sailem, J. Sero, R. Ardy, S. T. Wong, and C. Bakal, "How cells explore shape space: a quantitative statistical perspective of cellular morphogenesis," *Bioessays*, vol. 36, no. 12, pp. 1195–1203, 2014.

[18] D. L. Bodor, W. Pönisch, R. G. Endres, and E. K. Paluch, "Of cell shapes and motion: the physical basis of animal cell migration," *Developmental cell*, vol. 52, no. 5, pp. 550–562, 2020.

[19] Z. Pincus and J. Theriot, "Comparison of quantitative methods for cell-shape analysis," *Journal of microscopy*, vol. 227, no. 2, pp. 140–156, 2007.

[20] O. Tuzel, R. Subbarao, and P. Meer, "Simultaneous multiple 3d motion estimation via mode finding on lie groups," in *Tenth IEEE International Conference on Computer Vision, volume 1*, 2005, pp. 18–25.

[21] M. Ashizawa, H. Sasaki, T. Sakai, and M. Sugiyama, "Least-squares log-density gradient clustering for Riemannian manifolds," in *Proc. of the 20th International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2017.

[22] R. Subbarao and P. Meer, "Nonlinear mean shift over Riemannian manifolds," *International Journal of Computer Vision*, vol. 84, pp. 1–20, 2009.

[23] R. Caseiro, J. F. Henriques, P. Martins, and J. Batista, "Semi-intrinsic mean shift on Riemannian manifolds," in *Proc of ECCV, LNCS 7572,*, 2012, pp. 342–355.

[24] X. Deng, R. Sarkar, E. Labruyere, J.-C. Olivo-Marin, and A. Srivastava, "Characterizing cell populations using statistical shape modes," *ISBI*, 2021.

[25] A. Srivastava, I. Jermyn, and S. Joshi, "Riemannian analysis of probability density functions with applications in vision," in *IEEE Conference on computer Vision and Pattern Recognition (CVPR), Minneapolis, MN*, June 2007.

[26] S. H. Joshi, E. Klassen, A. Srivastava, and I. H. Jermyn, "A novel representation for riemannian analysis of elastic curves in $\mathbb{R}^n$," in *Proceedings of IEEE CVPR*, 2007, pp. 1–7.

[27] A. Srivastava, E. Klassen, S. H. Joshi, and I. H. Jermyn, "Shape analysis of elastic curves in euclidean spaces," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 33, no. 7, pp. 1415–1428, 2011.

[28] M. F. Abdelkader, W. Abd-Almageed, A. Srivastava, and R. Chellappa, "Gesture and action recognition via modeling trajectories on shape manifolds," *Computer Vision and Image Understanding Journal*, vol. 115, no. 3, pp. 439–455, 2011.

[29] C. Soto, D. Bryner, N. Neretti, and A. Srivastava, "Toward a three-dimensional chromosome shape alphabet," *J Comput Biol.*, vol. 28, no. 6, pp. 601–618, 2021.

[30] W. Liu, A. Srivastava, and J. Zhang, "A mathematical framework for protein structure comparison," *PLoS Computational Biology*, vol. 7, no. 2, p. e1001075, 2011.

[31] A. Srivastava, W. Wu, S. Kurtek, E. Klassen, and J. S. Marron, "Registration of functional data using fisher-rao metric," *arXiv*, vol. arXiv:1103.3817, 2011.