2022 IEEE 19th International Symposium on Biomedical Imaging (ISBI) | 978-1-6654-2923-8/22/$31.00 ©2022 IEEE | DOI: 10.1109/ISBI52829.2022.9761679

CHARACTERIZING CELL POPULATIONS USING STATISTICAL SHAPE MODES

Ximu Deng[1]*, Rituparna Sarkar*[2]*, Elisabeth Labruyere*[2]*, Jean-Christophe Olivo-Marin*[2] *& Anuj Srivastava*[1]

[1] Department of Statistics, Florida State University,Tallahassee, FL, USA
[2] Bioimage Analysis Unit, Institut Pasteur, Paris, France

ABSTRACT

We consider the problem of characterizing shape populations using highly frequent representative shapes. Framing such shapes as *statistical modes* – shapes that correspond to (significant) local maxima of the underlying *pdf*s – we develop a frequency-based, nonparametric approach for estimating sample modes. Using an elastic shape metric, we define ϵ-neighborhoods in the shape space and shortlist 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. The resulting modal set, in turn, helps characterize the shape population and performs better than the traditional cluster means. We demonstrate this framework using amoeba shapes from brightfield microscopy images and highlight its advantages over existing ideas.

Index Terms— Shape mode, Cell morphology, shape population, elastic shape analysis, Entamoeba histolytica.

1. INTRODUCTION

Cellular morphogenesis during migration is an exciting topic of study in various biological phenomena. Various intra- and extra-cellular physio-chemical changes induce morphological and positional changes in cells, reflecting the organism's ability to sustain itself in it's micro-environment. Cell migration, specifically amoeboid migration, is characterized by sequential protrusion and retraction of the cell membrane due to the restructuring of the cytoskeleton. This highly dynamic morphology introduces variation in the adopted cellular shapes during migration. Identifying the more prevalent shapes, a.k.a. *shape modes*, can aid in gaining insight into the cellular response to a particular micro-environment.

In the context of cell migration, shape analysis (using either static or dynamic shapes) has primarily been used for classifying migration patterns under different experimental conditions [1, 2, 3, 4, 5, 6]. A slightly different problem is discovering *dominant/frequent* shapes in a cell population and their variability across cell populations. This characterization can provide further insights into cell behavior and dynamics. However, past research has seldom focused on developing tools to identify such prevalent morphology.

Characterizing dominant statistical shapes in large data can be formalized in several ways. In Euclidean spaces, this can be done using an overall mean [7], or treat the population as a mixture of probability distributions (taken from a parametric family) and use an EM algorithm to estimate mixture parameters [8]. One can adapt these tools to the geometry of shape spaces using tangent space PCA, followed by Euclidean k-means clustering [9, 10, 11]. Due to the nonlinearity of shapes spaces and a preference for a nonparametric solution, we take a different approach. We consider the (unknown) underlying probability density function (pdf) on the shape space and seek its modes [12, 13, 14, 15]. These modes are defined as significant local maxima of the *pdf*, and one estimates *sample* modes using the observed shapes. We develop an efficient procedure that bypasses density estimation and seeks sample modes using a shape metric and ϵ-neighborhoods. This solution is similar in spirit to the k-mode clustering [16], originally presented for categorical data. Our approach has the following advantages: (1) It uses modes (instead of the means) as they are better shape representatives and simpler to compute; (2) It solves for "clustering" and modes simultaneously, rather than sequentially; (3) It does not assume any knowledge of k and is fully nonparametric; and, (4) It is much computationally efficient than the k-mean clustering.

2. DEFINING AND ESTIMATING SHAPE MODES

To set up a formal development, we define a shape space $\mathcal{S}$ and consider a *pdf* f on $\mathcal{S}$ denoting 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 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 [17, 18, 19] used for comparing cellular shapes.

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 imposing invariance to kinematics (rotation, translation, scaling, and re-parameterization of β). Let $[q]$ be the set of all rota-

Algorithm 1 Mode Estimation using a Discrete Setup

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:

$$\mathcal{N}_i = \{[q_j] : d_s([q_i], [q_j]) < \epsilon\}, i \neq j \quad (1)$$

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.

tions and re-parameterizations of a normalized SRVF q. The set of all shapes is denoted by $\mathcal{S} = \{[q]|q \in \mathbb{S}_\infty\}$. $\mathcal{S}$ is an infinite-dimensional, nonlinear space, and that limits our ability to perform traditional statistical analysis. Several tools have been developed in the past to study shapes as elements of $\mathcal{S}$. Given 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 (Karcher mean) and perform tangent PCA analysis for dimension reduction [20, 17]. These quantities – geodesics, shape distances, PCA, etc. – are invariant to rigid motions of curves and their parameterizations and are instrumental in removing cell kinematics from its morphology.

2.1. Nonparametric Mode Estimation - Mean-Shift

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 $\mathcal{S}$. One can choose a parametric or a nonparametric form of f for statistical analysis. Taking a nonparametric approach, one can estimate f as follows [21, 14]: For a Gaussian kernel, the kernel estimator takes the form: $\hat{f}([q]) \propto \sum_{i=1}^{n} e^{-d_s^2([q],[q_i])/\sigma^2}$. Since we are seeking the modes of f, we can ignore the normalization constant in $\hat{f}$ and seek its local maxima. These local extrema are located at the zeros of the gradient $\bigtriangledown_q \hat{f}$, and one can use a gradient search to find them. The full gradient-based algorithm is termed *nonlinear mean-shift* [14].

While this gradient-based search for modes of $\hat{f}$ is theoretically sound, it face some practical issues. One is the choice of the kernel K. The Gaussian kernel mentioned above is not always positive-definite on nonlinear manifolds. Also, the selection of the bandwidth σ is difficult. The larger problem of estimating a *pdf* on an infinite-dimensional, nonlinear manifold is problematic in itself.

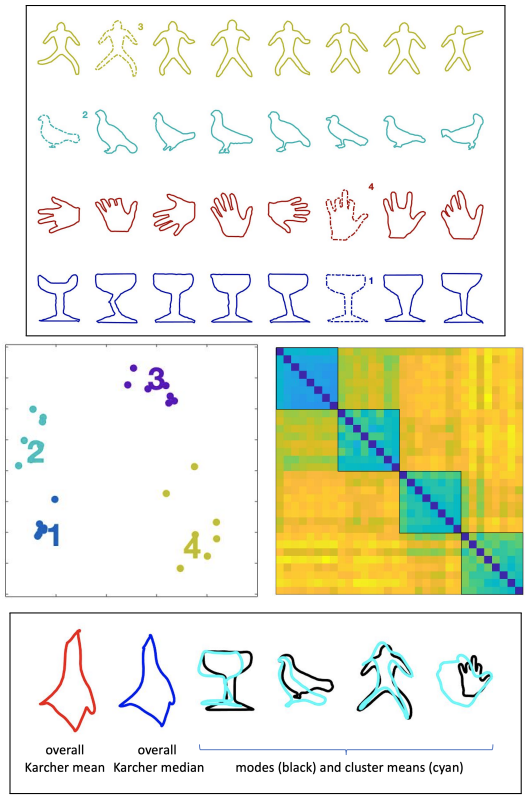
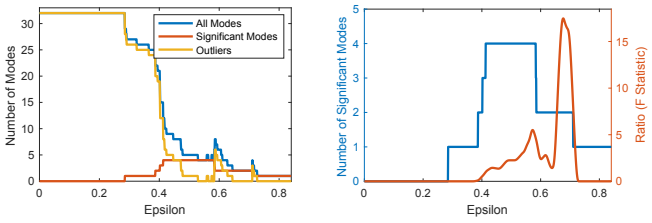


Fig. 1: Top: Shapes used in the experiment, with the estimated *modes* marked by numbers. Their cluster neighbors are drawn in the same color. Middle Left: 2D MDS plot for visualization of clusters with each shape shown as a point. Middle Right: Visualized pairwise distance matrix sorted by the clustering result from mode estimation. Blue denotes smaller distances and yellow denotes larger distances. Bottom: Overall Karcher mean (red), overall Karcher median (blue), modes (black) and cluster means (cyan).

2.2. Discretized Nonparametric Mode Estimation

We develop a discrete, frequency-based approach for finding shape modes while avoiding the onerous task of estimating the full *pdf*. This nonparametric approach is based only on the shape metric d_s mentioned earlier. We choose a scalar parameter, $\epsilon > 0$, that establishes the notion of a neighborhood under d_s in $\mathcal{S}$. For any curve $[q_i] \in \mathcal{S}$, any other shape $[q_j] \in \mathcal{S}$ is called its ϵ-neighbor if $d_s([q_i], [q_j]) < \epsilon$. The shapes with most neighbors are candidates for being the modes of $\hat{f}$. Given a shape dataset, and a fixed $\epsilon > 0$, the steps for estimating modes are summarized in Algorithm 1.

Selection of ϵ: While Algorithm 1 is straightforward, the specification of ϵ is non-trivial. Since the value of ϵ can significantly influence the results, the choice of ϵ is critical. In this paper, ϵ is computed using $\epsilon = 0.5\epsilon_M + 0.5\epsilon_F$. Here ϵ_M is chosen by maximizing the number of (*significant*) shape modes in the data. (We define a *significant* mode to be the one that has at least 2% neighbors, otherwise we label it an



(a) (b)

Fig. 2: 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. In (b), the blue curve shows number of *significant* modes and orange curve indicates *F-statistic* w.r.t ϵ.

outlier.) To define ϵ_F, we use the classical ANOVA (analysis of variance) but applied to the shape distances rather than shapes. We treat the pairwise distance $y_i = d_s([q_M], [q_i])$ between a mode and a shape as the response variable in ANOVA. Given an ϵ, the pairwise distances y_is between a mode and its ϵ-neighbors are considered as one group. We apply ANOVA on these y_i's to find the *F-statistics* as follows: Let g denote the number of clusters resulting from the chosen ϵ (using Algorithm 1). Set $s_k = \sum_{i=1}^{n_k} y_i$, where n_k is the size of the k^{th} cluster, and define:

$$SS_{total} = \sum_{i=1}^{n} y_i^2 - \frac{(\sum_{i=1}^{n} y_i)^2}{n}, SS_b = \sum_{k=1}^{g} \frac{s_k^2}{n_k} - \frac{(\sum_{k=1}^{g} s_k)^2}{n},$$

and set $SS_w = SS_{total} - SS_b$. Finally, compute: *F-statistics* $= \dfrac{SS_b/(g-1)}{SS_w/(n-g)}$. The *F-statistics* indicates how spread out the clusters are and ϵ_F is selected as the one that maximizes *F-statistics*. Using a fine grid on the interval $[0, \max d_s([q_i], [q_j])]$, we evaluate the potential ϵ_M, ϵ_F values and select the optima.

Illustrative Example: We demonstrate this approach with a simple experiment involving 32 shapes from four distinct classes - eight shapes in each class – as shown at the top part of Fig. 1. In Fig. 2, we plot the influence of ϵ on the number of modes and the *F-statistic*. Fig. 2(a) shows the values of all-modes, *significant* modes, and outliers versus ϵ. Fig. 2(b) displays the influence of ϵ on the number of significant modes and the *F-statistic*. The peak of the blue curve gives $\epsilon_M = 0.4131$ and the peak of the orange curve gives $\epsilon_F = 0.6773$, so $\epsilon = 0.5\epsilon_M + 0.5\epsilon_F = 0.5452$ and that yields **four distinct modes**. Fig. 1 (top) displays the mode shapes by labeling them as $1-4$ and corresponding cluster members in same color. The MDS plot (Fig. 1 (Middle-left)) shows shapes as planar points with colors denoting cluster memberships. Fig. 1 (middle-right) visualizes the pairwise distance matrix D between shapes arranged according to their clusters, with blue denoting smaller distances. The bottom row in Fig. 1 shows the overall Karcher mean (red),

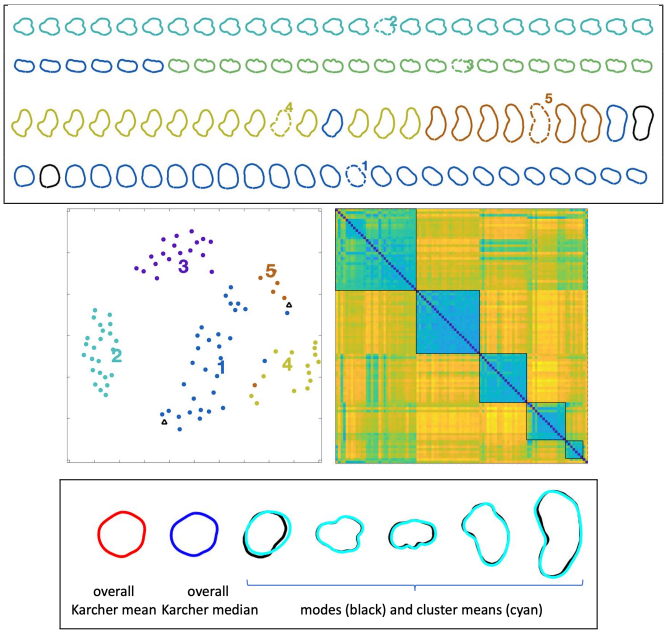
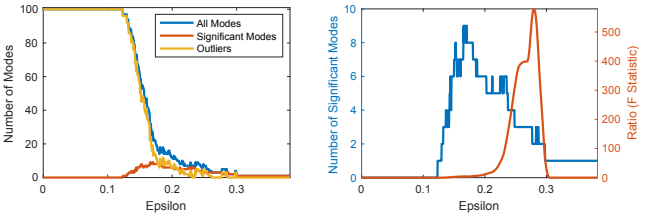


Fig. 3: Experiment 1 Results: layout, description is same as in Fig. 1.



(a) (b)

Fig. 4: Experiment 1 Results: layout, description is same as in Fig. 2.

Karcher median (blue), estimated modes (black) and cluster means (cyan). Due to significant variation within clusters, the cluster means lose critical shape features while the estimated modes retain characteristic features.

3. MODE ESTIMATION IN CELL POPULATIONS

In this section, we present some experimental results from our mode estimation on several sets of shapes of *Entamoeba histolytica* [3, 4].

Experiment 1: For this experiment, we select 100 cell shapes from four different cell migration sequences, *i.e.*, 25 shapes from each sequence, as shown in the top of Fig. 3. The plots for selecting ϵ are presented in Fig. 4, with peaks at $\epsilon_M = 0.1655$, $\epsilon_F = 0.2797$ and the optimal being $\epsilon = 0.2226$. For this ϵ, we discover **five modes** in the data. The shapes from each migration sequence contribute a mode, except the third sequence provides two modes. In Fig. 3, the MDS plot shows that the five clusters are well separated from each other. We also find two outliers in this data and they are

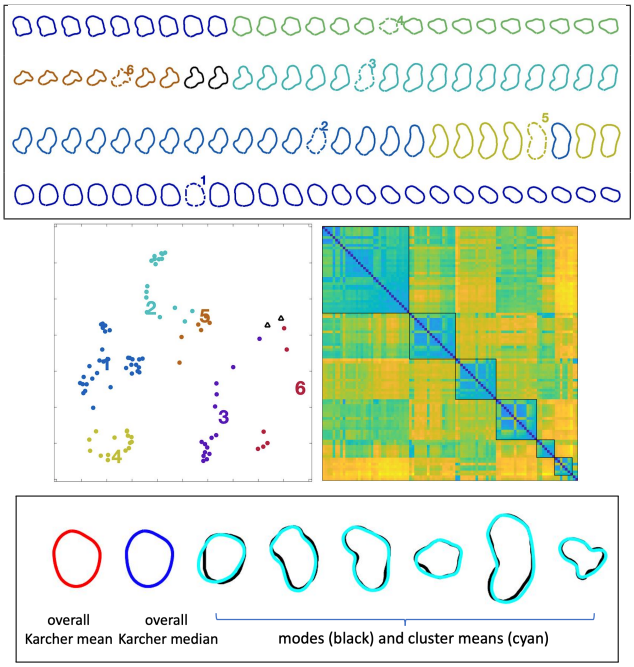


Fig. 5: Experiment 2 Results: layout, description is same as in Fig. 1.

marked with "△" in the MDS plot. (Keep in mind that the MDS plots are only for visualizing the clustering and are not always accurate in their depictions.)

Experiment 2: Once again we take 100 cell shapes from four different cell sequences (different from experiment 1) but the shapes are closer to each other this time. For this data, we obtain $\epsilon_M = 0.1655$, $\epsilon_F = 0.3111$ and the optimal $\epsilon = 0.2383$, resulting in **six modes**. The largest cluster contains all shapes from the fourth sequence and some from the first sequence. Fig. 5 (centre row) shows a 2D-MDS plot (left) and the distance matrix (right). Fig. 5 also shows the estimated modes and their improvements over cluster means.

Experiment 3: In this experiment, the dataset contains 100 cell shapes from 20 different cell sequences - five shapes from each sequence. Here we find $\epsilon_M = 0.1974$, $\epsilon_F = 0.3102$ and the optimal $\epsilon = 0.2538$. The data is well spread out and the algorithm finds **seven modes** and eight outliers. In the MDS plot in Fig. 6, data points seem to be distributed along a circle, pointing to the lack of a clear clustering pattern. Considering the relatively large variability in this dataset, the difference between the estimated modes and cluster means is much larger than earlier.

Table. 1 lists the sum of pairwise distances between a shape and its ϵ-neighbors. (We choose elements of the largest cluster in each experiment for this study.) This quantity is computed for the estimated mode and four randomly chosen shapes in that cluster. This quantity is the smallest for the mode shape, highlighting the centrality of modes. The last column in that table quantifies the agreement between our clustering and a traditional hierarchical clustering method. A

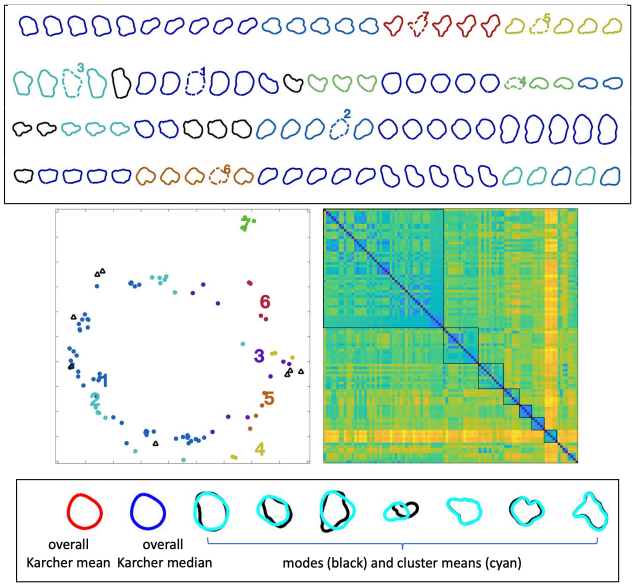


Fig. 6: Experiment 3 Results: layout, description is same as in Fig. 1.

Experiment	$[q_{M_1}]$	$[q_1]$	$[q_2]$	$[q_3]$	$[q_4]$	Matching Rate
1	**6.112**	6.152	6.186	6.302	6.353	95.9%
2	**6.583**	6.800	7.057	7.137	7.761	76.5%
3	**10.573**	11.708	11.842	11.903	12.066	53.3%

Table 1: Left part: Sum of pairwise distances from a shape to the remaining shapes in its cluster (for the largest cluster). Smallest values are for modes highlighting the **centrality of modes**. Last Column: The agreement between our clustering and hierarchical clustering.

value of 100% indicates a full agreement between the two approaches.

These experiments show that shape modes: (1) are superior representatives of shape populations than overall means/medians or cluster means, (2) provide a reasonable estimate of the number of clusters, (and 3) are obtained very efficiently despite shapes being infinite-dimensional and nonlinear. The main cost is in computing the pairwise distances between the given shapes. Thus, this approach provides a better solution than previous k-mean clustering or EM-based mixture solutions for shapes.

4. CONCLUSION

This paper introduces an efficient approach for characterizing cell populations using their modes. This nonparametric approach uses only pairwise distances and ϵ-neighborhoods, with ϵ determined automatically using a combination of ANOVA and modal distribution. The modal shapes are superior representations of shape populations compared with overall mean or cluster means.

5. COMPLIANCE WITH ETHICAL STANDARDS

This is a numerical simulation study for which no ethical approval was required.

Acknowledgements

This research was supported in part by the grants NSF DMS 1953087 and NIH R01 GM135927 to AS and in part by the France-BioImaging (FBI) infrastructure under Grant ANR-10-INBS-04, and the Program PIA INCEPTION under Grant ANR-16- CONV-0005.

6. REFERENCES

[1] 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.

[2] 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, pp. e1009237, 2021.

[3] 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.

[4] 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.

[5] 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.

[6] 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.

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

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

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

[10] Z. Yin, H. Sailem, J. Sero, R. Ardy, S. TC 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.

[11] Dani L Bodor, Wolfram Pönisch, Robert G Endres, and Ewa K Paluch, "Of cell shapes and motion: the physical basis of animal cell migration," *Developmental cell*, vol. 52, no. 5, pp. 550–562, 2020.

[12] 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.

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

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

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

[16] Z. Huang, "Extensions to the k-means algorithm for clustering large data sets with categorical values," *Data Mining and Knowledge Discovery*, vol. 2, pp. 283–304, 1998.

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

[18] 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.

[19] 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.

[20] 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.

[21] F Ferraty and P Vieu, *Nonparametric Functional Data Analysis: Theory and Practice*, Springer Series in Statistics. Springer New York, 2006.