35th Annual International Conference of the IEEE EMBS
Osaka, Japan, 3 - 7 July, 2013

A Novel Framework for Exploratory Analysis of Highly Variable Morphology of Migrating Epithelial Cells

Samuel D. R. Jefferyes[1], David B. A. Epstein[2], Anne Straube[3] and Nasir M. Rajpoot[1]

Abstract— **Migratory cells, for example human retinal epithelial (RPE) cells, exhibit highly variable morphology. This makes it difficult to use traditional methods, such as the landmark based Procrustes analysis or feature based analysis, to quantitatively represent their shapes. We propose a novel framework to generate a low-dimensional representation of highly variable cell shapes. The framework lends itself readily to efficient exploratory analysis of a given cell shape dataset in order to visualise morphological trends in the data and reveal the intrinsic structure of various morphology-based cell phenotypes in the data. Preliminary results show that the framework is effective in revealing consistent morphological phenotypes.**

I. INTRODUCTION

Directed cell migration is important for many physiological processes, including embryonic development [1] and wound healing [2]. Deregulated cell migration causes human disease, such as tumour metastasis in cancer [3]. In this work we study morphological variations in time-lapse images of human epithelial cells undergoing random migration in culture (Fig. 1). Cell shapes reflect different modes of directional change and the cyclic formation/retraction of cell tails, a mechanism that controls migration directionality in epithelial cells [4]. Shape variation has thus far been restricted to the analysis of relatively small cells or organelles with limited morphological variability [5], [6], [7]. Two common algorithmic approaches are popular in morphological analysis:

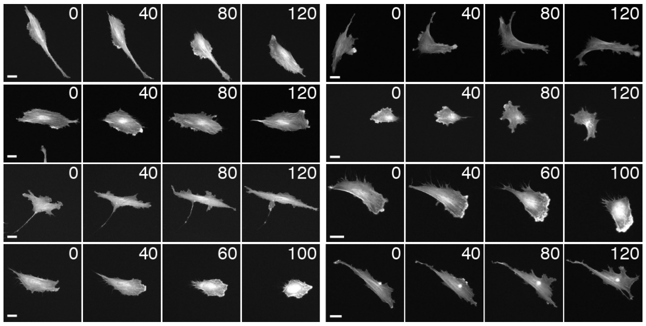

Fig. 1. Representative frames of migrating RPE1 cells expressing mGFP-LifeAct to mark actin. Scale bars are 20μm and relative time is indicated in minutes.

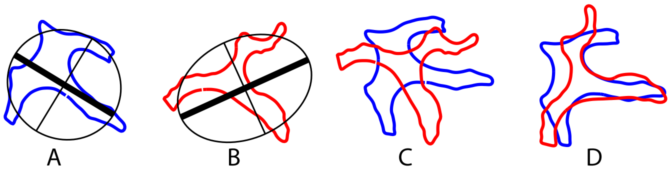

Fig. 2. This figure illustrates the difficulties faced when mutually aligning complex shapes along intrinsic axes. The curves labelled A & B show cell contours and their best-fit ellipses with thick major axis. C shows the result of aligning the two major axes (a common approach). D shows a more suitable alignment of the two shapes.

1) *Standard form alignment.* This can be through landmark registration [5], which is limited to data with consistent landmarks (not present in our RPE cells); or principal axes alignment [8] where a small difference in shape can lead to a discontinuity in the axis of alignment (Fig. 2).

2) *Measurable shape feature representation.* The strategy of measuring a finite set of selected shape feature vectors (for example [7]) will only ever represent limited degrees of variation. While this technique can be appropriate in some applications, it is often insufficient to reliably capture the structure of highly variable cell morphology data.

This has led us to develop a quantitative morphological descriptor that does not impose an importance on any individual morphological properties such as size or roundness, but attempts to discern the prominent areas of shape variation present within any given dataset of cell images. Rajpoot & Arif [9] demonstrate the effectiveness of an unsupervised manifold learning technique called diffusion maps [10] in classifying images of objects by shape similarity. The framework creates a low dimensional representation of shape space and extracts the intrinsic variability within a dataset of shapes.

In this paper, we present a novel framework for exploratory analysis of intrinsic morphological variations employing an elastic metric based geodesic distance for shape similarity comparison. Fig. 3 illustrates the major building blocks of our representation framework. First the cell contours are segmented from the images. Then the contours are converted to a shape descriptor representation that facilitates a rapid computation of a low-dimensional representation of the set of shapes through a manifold learning algorithm. This allows for visualisation and quantitative analysis of the morphological behaviour of individual cells as well as trends in cell populations.

[1] S. Jefferyes and N. Rajpoot are with the Computational Biology and Bioimaging Group, Department of Computer Science, University of Warwick, United Kingdom; Corresponding authors: {s.d.r.jefferyes, n.m.rajpoot}@warwick.ac.uk

[2] D. Epstein is with the Warwick Mathematics Institute, University of Warwick, United Kingdom

[3] A. Straube is with the Centre for Mechanochemical Cell Biology, Division of Biomedical Cell Biology, University of Warwick, United Kingdom

978-1-4577-0216-7/13/$26.00 ©2013 IEEE

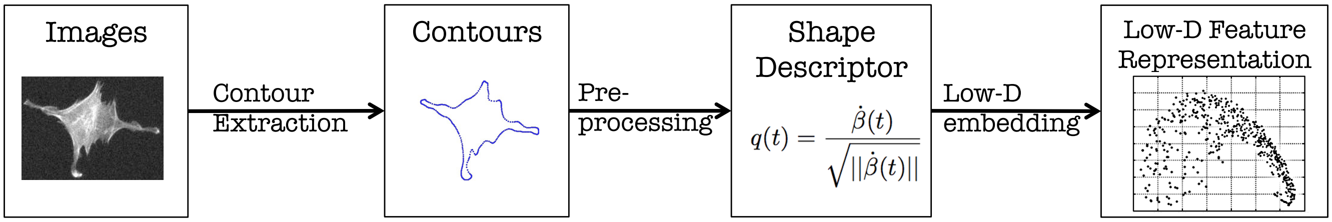
Images
Contour Extraction
Contours
Pre-processing
Shape Descriptor
$q(t) = \frac{\dot\beta(t)}{\sqrt{||\dot\beta(t)||}}$
Low-D embedding
Low-D Feature Representation

Fig. 3. A flow diagram illustrating the algorithm developed for generation of a low dimensional representation of cell shapes. We make repeated use of this algorithm for quantitative cell shape analysis on images of migrating cells.

II. MATERIALS AND METHODS

A. Cell Culture and Live Cell Imaging

Human retinal pigment epithelial cells (RPE1) expressing mGFP-LifeAct were seeded in a glass-bottom chamber coated with 10 μg/ml fibronectin. Cells were imaged every 5 minutes for 12 hours in a stage top incubator maintaining 37°C and 5% CO_2 on a Deltavision system using a 20× 0.85NA oil immersion objective, a GFP filter set, and a CoolsnapHQ camera under control of SoftWorx.

B. Contour Extraction

Cell outlines were extracted for every time point using Quimp10 [11]. We represent each cell shape as a circular sequence f of N uniformly spaced points in the plane $f = \{(x_1, y_1), (x_2, y_2), \dots, (x_N, y_N)\}$. Note that this sequence f represents a closed curve with equidistant points, i.e. $||(x_N, y_N) - (x_1, y_1)|| = ||(x_i, y_i) - (x_{i+1}, y_{i+1})||$ for any $i \in \{1, \dots N - 1\}$.

C. Shape Similarity Measure using the Elastic Metric

The success of our exploratory analysis framework depends largely on the selection of an appropriate similarity measure. This is crucial to the process because it is through the similarity measure that the algorithm learns the intrinsic geometric structure of the data. One major challenge relates to invariance under transformations; the measure of difference between two curves must be invariant with respect to change of orientation and change of parameterisation of one curve, while these are unchanged for the other. A separate, and only subtly different point, is that any two curves need to be compared with appropriate relative alignment and parameterisation. This is discussed in detail in [12].

Our similarity measure employs the notion of a geodesic distance (as described in [13]) that corresponds to the length of a path through shape space minimising deformation between the target shapes. This deformation is measured through the use of an elastic metric [14] which quantifies the bending and stretching required to deform between shapes. Given $q_0, q_1 \in \mathcal{C}$, where $\mathcal{C}$ is the Riemannian manifold representing the space of curves in the plane, let $\alpha : [0, 1] \to \mathcal{C}$ be a parameterised path with $\alpha(0) = q_0$ and $\alpha(1) = q_1$. Then we can define the length of path α to be $L(\alpha) = \int_0^1 \sqrt{\langle\dot\alpha(t), \dot\alpha(t)\rangle}dt$, according to Elastic metric $\langle\cdot,\cdot\rangle$, and we can define the distance between q_0 and q_1 as

$$d_c(q_0, q_1) = \inf_{\alpha} L(\alpha). \qquad (1)$$

where α ranges over all paths $\alpha : [0, 1] \to \mathcal{C}$ with $\alpha(0) = q_0$ and $\alpha(1) = q_1$.

In order to introduce necessary invariance to in-plane transformations, we look at shape space ($\mathcal{S}$) as a quotient of the space of curves by the groups of reparameterisations (Γ) and rotations in the plane ($SO(2)$) i.e. $\mathcal{S} = \mathcal{C}/(\Gamma \times SO(2))$. The geodesic distance between two closed curves q_0 and q_1 is then defined as,

$$d_{\mathcal{S}}([q_0], [q_1]) = \inf_{\{(\gamma,\mathcal{O})\in\Gamma\times SO(2)\}} d_c(q_0, \mathcal{O}(q_1 \circ \gamma)\sqrt{\dot\gamma}). \qquad (2)$$

Note that $\mathcal{O}(q \circ \gamma)\sqrt{\dot\gamma}$ is the operation of $(\gamma, \mathcal{O})$ on q in the Square-Root Velocity representation [13].
Returning to our discretised contours, if the geodesic distance between contours f_j and f_k is $d(f_j, f_k)$, we define the shape similarity measure to be,

$$w(f_j, f_k) = \exp\left(\frac{d_{\mathcal{S}}(f_j, f_k)^2}{\sigma^2}\right) \qquad (3)$$

To the best of our knowledge, bandwidth determination is still an open problem. We make use of reverse soft K-nearest neighbour density estimation [15] to determine σ. Upon perturbation of σ by up to 15%, clustering agreement with unperturbed results remains high (Rand index >0.93).

D. Diffusion Maps

The diffusion maps framework [10] is a non-linear dimensionality reduction technique that generates a low-dimensional coordinate representation of data. Similar data points in the high-dimensional shape space are represented by new low-dimensional points that are close; dissimilar data points are represented by new low-dimensional points that are far apart.

To perform a diffusion maps based low-dimensional embedding of n contours, $\{f_i\}$ where $1 \leq i \leq n$, one constructs an $n \times n$ matrix P with its (j, k)th entry given as follows,

$$p_{jk} = \frac{w(f_j, f_k)}{\sum_i w(f_j, f_i)} \qquad (4)$$

where $w(\cdot, \cdot)$ is the chosen shape similarity measure. This matrix P, can be thought of as a Markov transition matrix (where similarity is analogous to diffusion distance). Then we perform eigen-decomposition upon P, and we know by the Perron-Frobenius theorem that P has exactly one eigenvalue equal to 1 and all other eigenvalues have strictly smaller magnitude. So (by reordering if necessary) let $1 = \lambda_0 > |\lambda_1| \geq |\lambda_2| \geq \ldots \geq |\lambda_{n-1}|$ be the set of eigenvalues, and $\{\psi_i | i = 0, \ldots, n-1\}$ be the set of corresponding n-dimensional eigenvectors. Then, if $\psi_i^{(j)}$ is the jth component of the eigenvector ψ_i, we construct a lower dimensional representation of contour f_j as

$$\varphi_j = (\lambda_1^t \psi_1^{(j)}, \lambda_2^t \psi_2^{(j)}, \ldots, \lambda_\rho^t \psi_\rho^{(j)}) \tag{5}$$

where $\rho \ll n$ is our choice of dimension for the embedding, and t denotes time in the Markovian sense (we chose $t = 1$ in our analysis, as we are interested in local geometric properties of shape space). Note that ρ is chosen to be much lower than the dimensionality of the original data, and hence φ_j is a low dimensional embedding of the contours. In a similar fashion to other dimensionality reduction techniques, $|\lambda_i|$ reflects the proportion of the overall variance of the dataset that is accounted for in eigenvector ψ_i. Hence ρ can be chosen large enough to give the desired accuracy.

E. Morphological Feature based Analysis

To assess the performance of the proposed framework, we compared it with the morphological feature based analysis employed in [6] using the morphological features offered by CellCognition [7], a tool for shape and texture based morphological analysis of cells. Each shape contour was converted into a binary mask. Then the following 8 shape features were computed: area, circularity, dist_max, dist_min, dist_ratio, foreground irregularity, background irregularity, and perimeter (see [7] for more details). A z-score normalisation and principal component analysis were performed on the feature vectors in order to obtain the low-dimensional representation of cell shapes.

F. Hierarchical Clustering

Hard clustering is not the most appropriate way of examining the structure of the low-dimensional representation of cell shapes, since often the dataset lies as a continuous point cloud and not as distinct clusters. For this reason, we do not greatly concern ourselves with achieving high cluster validity. However, clustering does allow us to explore the groups of high-dimensional data (contours) that are embedded to different parts of the low-dimensional point cloud in an unsupervised manner. Then the success of the low-dimensional representation can be judged by how well the cluster validity is preserved in the low-dimensional space, i.e. how well each group of contours represents a distinct morphological phenotype. We used hierarchical clustering using Ward's minimum variance method [16].

III. RESULTS AND DISCUSSION

In this section, we present experimental results for 500 cell contours extracted from 25-frame time-lapse image sequences of migrating RPE1 cells. Fig 4(A) contains a scatterplot of the top three principal components of the shape feature vectors computed for each shape (see Sec. II.E), while Fig 4(B) contains a scatterplot of the top three embedding coordinates of the dataset using the diffusion maps framework (see Sec. II.C & II.D). We performed hierarchical clustering [16] in order to compare the potential for exploratory analysis of the two low-dimensional representations. The scatter plots in Fig. 4 are coloured according to the partitioning into 6 clusters. 8 contours were randomly chosen from each cluster and are displayed below the hierarchical dendrogram with the corresponding cluster colour. There is no *ground truth* for the membership of clusters, and so there is no objective criterion to say that one clustering is better than another. Instead we use a frankly subjective approach, examining sample contours with cluster labels and assessing the inter- and intra- cluster shape similarity.

Results shown for feature based analysis in Fig. 4(A) show a successful separation of simple round shapes. However, the distribution of different complex shapes is not captured effectively. For instance, cluster 6 of Fig. 4(A) (in red) contains a wide variety of morphological phenotypes. Thus this method is most suitable for the classification of round shapes as required for the annotation of mitotic stages ([6] and [7]).

Fig. 4(B) displays results of hierarchical clustering using our proposed new method. This gives a good separation of morphological phenotypes. It can be observed from this result that (a) each cluster seems to contain contours of a particular phenotype and (b) the perceivable *average shape* from each cluster seems different to the others. When examining the higher levels of the dendrogram, we again see reasonable agreement. For example, cluster 4 in Fig. 4(B) is arguably more similar to clusters 1–3 (its *cousins* in the hierarchy) than clusters 5–6. The proposed framework has been successful at distributing the points (each point corresponding to a contour) so as to reflect the shape similarity through Euclidean distance in our low dimensional space of coordinates.

IV. CONCLUSIONS

The main contribution of this paper is the presentation of a framework for exploratory analysis of morphology based phenotypes of cells with highly variable shapes. The proposed approach is landmark free, is completely unsupervised, does not require computation of any explicit morphological feature measurements, and captures the intrinsic non-linear structure of the high-dimensional shape space of highly variable morphologies in our RPE cell shape data. It is computationally expensive to compute distances within the current framework, for all pairs of a large dataset. In the future we will instead use our own rapidly computed distance and similarity measures. We believe this will be useful for answering several biological and clinical questions.

ACKNOWLEDGEMENTS

We thank Sarah Cosgriff for the extraction of cell outlines and preparation of data for shape analysis, Adnan Khan for discussion on clustering, and Ben Fitton for help with

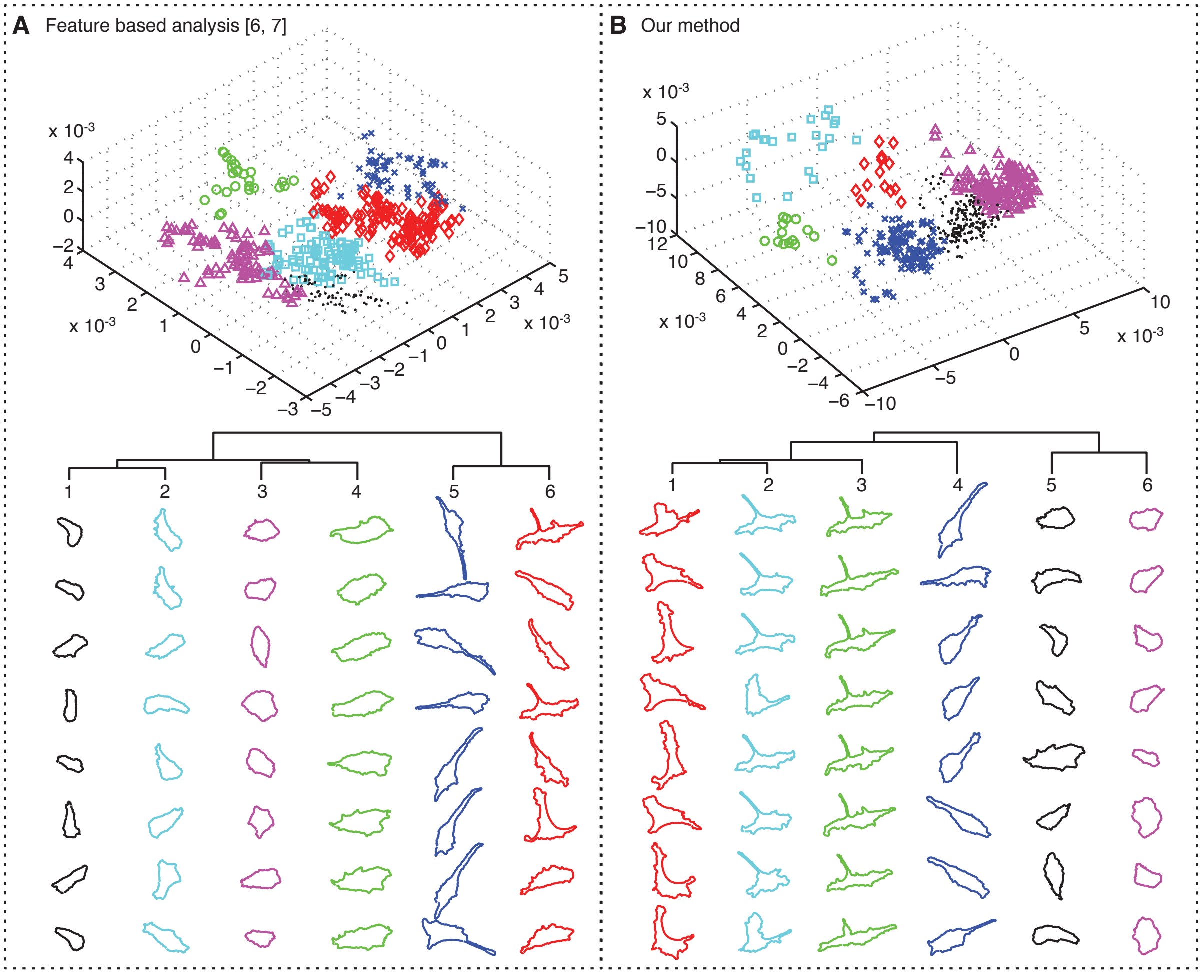

Fig. 4. The top 2 scatterplots show clustering of 500 RPE1 cell contours using (A) principal component analysis performed on shape features and (B) the geodesic distance with the diffusion framework. The axes in (A) are the top 3 principal components. The axes in (B) indicate distances after diffusion. Relative Euclidean distance should reflect shape similarity between the represented data points. Hierarchical clustering was performed on each set of embedded points. After constructing the clustering hierarchy, we chose the level which gave 6 clusters, and assigned a colour to each cluster. The points in the scatter plots are coloured according to their cluster. 8 cell contours were randomly selected from each cluster and these are shown in columns and with corresponding cluster colour. Dendrograms illustrate the cluster linkage at higher levels.

figure preparation. S. J. is supported by the BBSRC funded Warwick Systems Biology Doctoral Training Centre and A. S. is supported by a Marie Curie Cancer Care grant.

References

[1] R. Keller. Cell migration during gastrulation. Current Opinion in Cell Biology, 17(5):533-41, 2005.

[2] W.S. Krawczyk. A pattern of epidermal cell migration during wound healing. The Journal of Cell Biology, 49(2):247-63, 1971.

[3] W. Wang, S. Goswami, E. Sahai, J.B. Wyckoff, J.E. Segall, and J.S. Condeelis. Tumor cells caught in the act of invading: their strategy for enhanced cell motility. Trends in Cell Biology, 15(3):138-45, 2005.

[4] U. Theisen, E. Straube and A. Straube, Directional Persistence of Migrating Cells Requires Kif1C-Mediated Stabilization of Trailing Adhesions. Dev. Cell, vol. 23 (6) pp. 1153-1166, 2012.

[5] K. Keren, Z. Pincus, G. M. Allen, E. L. Barnhart, G. Marriott, A. Mogilner, and J. Theriot, Mechanism of shape determination in motile cells, Nature, vol. 453, no. 7194, pp. 475-80, 2008.

[6] Q. Zhong, A. G. Busetto, J. P. Fededa, J. M. Buhmann, and D. W. Gerlich, Unsupervised modeling of cell morphology dynamics for time-lapse microscopy., Nature Methods, 9(7):711-3, 2012.

[7] M. Held, M. H. Schmitz, B. Fischer, T. Walter, B. Neumann, M. H. Olma, M. Peter, J. Ellenberg, and D. W. Gerlich, CellCognition: time-resolved phenotype annotation in high-throughput live cell imaging., Nature Methods, 7(9):747-54, 2010.

[8] Z. Pincus and J. Theriot, Comparison of quantitative methods for cell-shape analysis., Journal of Microscopy, 22(2):140-56, 2007.

[9] N. Rajpoot and M. Arif, Unsupervised shape clustering using diffusion maps, Annals of the BMVA, 2008(5):1-17, 2008.

[10] R. Coifman and S. Lafon, Diffusion maps. Applied and Computational Harmonic Analysis, 21(1):5-30, 2006.

[11] R.A. Tyson, D.B.A. Epstein, K.I. Anderson, and T. Bretschneider, High resolution tracking of cell membrane dynamics in moving cells: An electrifying approach. Math. Model. Nat. Phenom., 5(1):34-55, 2010.

[12] X.R. Lele and J. T. Richtsmeier. An invariant approach to statistical analysis of shapes. Chapman and Hall/CRC, 2001.

[13] 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, 33(7):1415-1428, 2011.

[14] W. Mio, A. Srivastava, and S.H. Joshi, On shape of plane elastic curves, Intl. J. Computer Vision, 73(3):307-324, 2007.

[15] O. Kursun, Spectral clustering with reverse soft K-nearest neighbor density estimation, Proceedings International Joint Conference onNeural Networks (IJCNN), pp. 1-8, 2010.

[16] J.H. Ward Jr., Hierarchical grouping to optimize an objective function, Journal of the American Statistical Association, 58(301):236-244, 1963.