Phenotyping Cellular Motion

Felix Yuran Zhou

Balliol College
University of Oxford

A thesis submitted for the degree of
Doctor of Philosophy

Michaelmas 2017

Abstract

In the development of multicellular organisms, tissue development and homeostasis require coordinated cellular motion. For example, in conditions such as wound healing, immune and epithelial cells need to proliferate and migrate. Deregulation of key signalling pathways in pathological conditions causes alterations in cellular motion properties that are critical for disease development and progression, in cancer it leads to invasion and metastasis. Consequently there is strong interest in identifying factors, including drugs that affect the motion and interactions of cells in disease using experimental models suitable for high-content screening. There are two main modes of cell migration; individual and collective migration. Currently analysis tools for robust, sensitive and comprehensive motion characterisation in varying experimental conditions for large extended timelapse acquisitions that jointly considers both modes are limited.

We have developed a systematic motion analysis framework, Motion Sensing Superpixels (MOSES) to quantitatively capture cellular motion in timelapse microscopy videos suitable for high-content screening. MOSES builds upon established computer vision approaches to deliver a minimal parameter, robust algorithm that can i) extract reliable phenomena-relevant motion metrics, ii) discover spatiotemporal salient motion patterns and iii) facilitate unbiased analysis with little prior knowledge through unique motion 'signatures'. The framework was validated by application to numerous datasets including YouTube videos, zebrafish immunosurveillance and *Drosophila* embryo development. We demonstrate two extended applications; the analysis of interactions between two epithelial populations in 2D culture using cell lines of the squamous and columnar epithelia from human normal esophagus, Barrett's esophagus and esophageal adenocarcinoma and the automatic monitoring of 3D organoid culture growth captured through label-free phase contrast microscopy. MOSES found unique boundary formation between squamous and columnar cells and could measure subtle changes in boundary formation due to external stimuli. MOSES automatically segments the motion and shape of multiple organoids even if present in the same field of view. Automated analysis of intestinal organoid branching following treatment agrees with independent RNA-seq results.

Phenotyping Cellular Motion

Felix Yuran Zhou

Balliol College

University of Oxford

A thesis submitted for the degree of

Doctor of Philosophy

Michaelmas 2017

Acknowledgements

First and foremost, I would like to express my gratitude to my two DPhil supervisors, Profs. Jens Rittscher and Xin Lu. Over the past three and a half years they have guided, challenged and inspired me. I thank Xin for daring to hire an engineer with little knowledge of biology and providing me with numerous challenging projects and Jens for giving me the freedom to explore my computational ideas. I also wish to acknowledge my funding sources; an EPSRC Life Sciences Interface Doctoral Training Centre stipend and the Ludwig Institute for Cancer Research without which this DPhil would not be possible.

I extend my gratitude to my examiners Profs. Shankar Srinivas and Patrick Bouthemy for their time and effort in examining me.

A massive thanks to all of my past and present Ludwig colleagues for providing a supportive, fun and educational environment to work in. Thanks to Ludovico Buti who supervised me as a rotation student and together with Richard Owen and Zinaida Dedeić provided the initial videos for prototyping MOSES, David Severson for RNA-seq assistance, Mike White, Jingyi Ma for histology slides and Xiao Qin for organoid videos. Special thanks to Carlos Ruiz-Puig who has worked closely with me over the years and often has had to do such thankless jobs as laborious manual annotation. I also thank colleagues outside the Ludwig. A huge thanks to Weidi Xie who introduced me initially to deep learning and for many stimulating machine learning discussions and Thomas Nketia for cell tracking discussions.

I gratefully acknowledge my many academic and clinical collaborators who have either provided me data or invaluable advice. Thanks to Svanhild Nornes and Prof. Mads Gyrd-Hansen for the zebrafish surveillance videos, Chandan Seth Nanda, Ruud Van Stiphout and Prof. Sebastian Nijman for work on phenotyping cell states from high-content microscopy cell images, Prof. Robert Goldin for his histology expertise, Dong Xin, Prof. Yanning Gao, Dr Xiao-Guang Ni for work investigating EBV in NPC and their kind hospitality in Beijing.

I must thank Profs. Gil McVean and Alison Noble for their sound advice in my DPhil confirmation without which this thesis would not have been born.

Last I thank my family; my parents, Jian Guo Zhou and Xieping Li for nurturing my passion for mathematics and science from a young age, my sister Clare Zhou and all my friends whom have shared in my excitements and commiserations.

I am indebted to you all without which this DPhil could not have been achieved. Thank you all very much.

Declaration

All the material presented in this thesis is the result of my own work, except where specifically indicated here and in the corresponding sections:

1. In chapter 4, the zebrafish immunosurveillance videos were acquired by Svanhild Nornes in collaboration with Prof. Mads Gyrd-Hansen.

2. In chapter 5, all the wetlab work including development of the experimental *in-vitro* assay, fluorescent staining and timelapse microscopy acquisition was carried out by Carlos Ruiz-Puig. Carlos also annotated the gap closure frames.

3. In chapter 6, all the wetlab work including the organoid culture, timelapse acquisition and RNA-seq preparation was carried out by Xiao Qin. Xiao Qin and Carlos Ruiz-Puig manually counted the organoids. Carlos Ruiz-Puig manually annotated the organoid branching. RNA-seq was carried out by the Wellcome Trust Centre for Human Genetics.

Contributions by the aforementioned colleagues are indicated in the text of the thesis at the relevant points. None of the work presented in this thesis has been accepted or submitted for any degree, diploma or certificate or other qualification in this University or elsewhere. This thesis is 42,658 words and does not exceed the 50,000 words as recommended by the Medical Sciences Division.

Felix Yuran Zhou, 2017

Abstract

In the development of multicellular organisms, tissue development and homeostasis require coordinated cellular motion. For example, in conditions such as wound healing, immune and epithelial cells need to proliferate and migrate. Deregulation of key signalling pathways in pathological conditions causes alterations in cellular motion properties that are critical for disease development and progression, in cancer it leads to invasion and metastasis. Consequently there is strong interest in identifying factors, including drugs that affect the motion and interactions of cells in disease using experimental models suitable for high-content screening. There are two main modes of cell migration; individual and collective migration. Currently analysis tools for robust, sensitive and comprehensive motion characterisation in varying experimental conditions for large extended timelapse acquisitions that jointly considers both modes are limited.

We have developed a systematic motion analysis framework, Motion Sensing Superpixels (MOSES) to quantitatively capture cellular motion in timelapse microscopy videos suitable for high-content screening. MOSES builds upon established computer vision approaches to deliver a minimal parameter, robust algorithm that can i) extract reliable phenomena-relevant motion metrics, ii) discover spatiotemporal salient motion patterns and iii) facilitate unbiased analysis with little prior knowledge through unique motion 'signatures'. The framework was validated by application to numerous datasets including YouTube videos, zebrafish immunosurveillance and *Drosophila* embryo development. We demonstrate two extended applications; the analysis of interactions between two epithelial populations in 2D culture using cell lines of the squamous and columnar epithelia from human normal esophagus, Barrett's esophagus and esophageal adenocarcinoma and the automatic monitoring of 3D organoid culture growth captured through label-free phase contrast microscopy. MOSES found unique boundary formation between squamous and columnar cells and could measure subtle changes in boundary formation due to external stimuli. MOSES automatically segments the motion and shape of multiple organoids even if present in the same field of view. Automated analysis of intestinal organoid branching following treatment agrees with independent RNA-seq results.

Contents

List of Figures

List of Tables

List of Abbreviations

1d, 2d, nd . . One-, two-dimensional or n-dimensional, referring to dimensions of abstract feature vectors.

1D, 2D, nD . . One-, two-dimensional or n-dimensional, referring to spatial dimensions in an image or geometrically.

1.s.f, 2.s.f, 3.s.f 1, 2 or 3 significant figures

BE . . . . . Barrett's Esophagus

BoW . . . . . Bag of Words

CNN . . . . . Convolutional Neural Network

CPM . . . . . Counts Per Million

CRF . . . . . Conditional Random Field

CT . . . . . Computed Tomography

DHO . . . . . Damped Harmonic Oscillator

DIC . . . . . Differential Interference Contrast microscopy

DNA . . . . . Deoxyribonucleic acid

DOS . . . . . Density of States

EAC . . . . . Esophageal Adenocarcinoma

ECM . . . . . Extracellular Matrix

EGF . . . . . Epidermal Growth Factor Receptor

EGFR . . . . . Epidermal Growth Factor

Expt . . . . . Short for experiment

FBS . . . . . Fetal Bovine Serum

FTLE . . . . . Finite Time Lyapunov Exponent

FV . . . . . Fisher Vector encoding

GBH . . . . . Gradient Boundary Histogram

GCN . . . . . Graph Convolutional Network

GMM . . . . . Gaussian Mixture Model

GSK . . . . . Refers to GSK-3, a protein kinase that mediates the addition of phosphate molecules onto serine and threonine amino acid residues

h, hr . . . . . Short for hour or hours

HDP-HMM . Hierarchical Dirichlet Process Hidden Markov Model

HMM . . . . . Hidden Markov Model

HoF . . . . . Histogram of Oriented Flow

HoG . . . . . Histogram of Oriented Gradients

IDT . . . . . Improved Dense Trajectories

KLT . . . . . Kanade-Lucas Tracker

kNN . . . . . k-Nearest Neighbours

KO . . . . . Refers to knock-out condition

LCS . . . . . Lagrangian Coherent Structures

LoG . . . . . Laplacian of Gaussian

LSTM . . . . . Long Short-Term Memory

MAD . . . . . Mean Absolute Deviation

MAE . . . . . Mean Absolute Error

MBH . . . . . Motion Boundary Histogram

MDS . . . . . Multidimensional Scaling

MOSES . . . . Motion Sensing Superpixels

MRF . . . . . Markov Random Field

MSD . . . . . Mean Squared Displacement

MSE . . . . . Mean Squared Error

PCA . . . . . Principal Components Analysis

PDE . . . . . Partial Differential Equation

PIV . . . . . Particle Image Velocimetry

RAG . . . . . Region Adjacency Graph

RGB . . . . . Refers to the colour model that represent images with red, green, and blue channels

RMSD . . . . . Root Mean Squared Displacement

RNA . . . . . Ribonucleic acid

RNN . . . . . Recurrent Neural Network

SCJ . . . . . Squamous Columnar Junction

SEM . . . . . Standard Error of the Mean

SIFT . . . . . Scale-Invariant Feature Transform

SVM . . . . . Support Vector Machine

TCC . . . . . Track Cross Correlation

TSN . . . . . Temporal Segment Network

t-SNE, TSNE t-distributed stochastic neighbour embedding

VC . . . . . Refers to combined valproic acid + CHIR99021 (a GSK inhibitor) treatment

VCC . . . . . Velocity Cross Correlation

VPA . . . . . Valproic Acid

WT . . . . . Refers to wild-type condition

1 General Introduction

Contents

1.1 Overview and Significance

1.1.1 Biological Motion and Disease

The movement of cells is fundamental to the development and maintenance of multicellular organisms and plays a critical role in numerous cellular processes. Complex patterns of cell migration are essential for proper tissue and organ formation during embryogenesis (Montell 2008). Wound healing involves the coordinated migration of multiple cell types, (Ridley et al. 2003). The migration of leukocytes into lymph nodes and inflamed tissue is required for the development of immune responses (Ridley et al. 2003; Friedl and Weigelin 2008). When defective, cell

migration can cause disease; one of the hallmarks of cancer is invasion and metastasis (Hanahan and Weinberg 2011). There are two principal mechanisms cells migrate *in-vivo*, either moving as single cells or collectively where cells remain connected to each other as they move resulting in migrating cohorts and varying degrees of tissue organization (Vaughan and Trinkaus 1966; Friedl 2004; Montell 2008). The migration of single cells is the best-studied mechanism of cell movement *in-vitro* and is known to contribute to many physiological motility processes in-vivo, such as development, immune surveillance and cancer metastasis (Ridley et al. 2003; Friedl and Weigelin 2008). Collective migration in particular drives the formation of many complex tissues and organs such as that in embryogenesis. Therefore there is strong biological and clinical interest and need to study precise cellular motion characteristics in a tissue-relevant context and to identify factors, including drugs, that affect the motion and interactions of cells using experimental models suitable for high-content screening.

1.1.2 Experimental Models to Investigate Biological Motion

To systematically investigate cellular motion in the laboratory, cell culture experimental models are used. 2D *in-vitro* cell culture is commonly used with cells plated sparsely to investigate the individual cellular motion and changes to morphology or as a confluent monolayer for example with epithelial cells to investigate the effect of collective motion. Geometric patterning such as wound healing assays where a predefined gap is introduced in the monolayer is often used to measure the collective migration speed of the cells based on how fast the gap is closed, (Liang et al. 2007; Masuzzo et al. 2016). More complex spatial control using micropatterning techniques can be used to scale up the basic wound healing assay or to investigate the interactions between two or more populations of cells. The ease of scaling, cell manipulation and imaging with 2D culture has made this a widely used platform for high content image based analysis and screening. However there are several key limitations for translating the resultant findings using 2D cell culture systems to

the clinic. In 2D cell culture, cells are grown on flat dishes made of polystyrene plastic that is very stiff and unnatural. The cells adhere and spread on this plastic surface and form unnatural cell attachments to proteins that are deposited and denatured on this synthetic surface. In 3D *in-vivo*, the behaviour of cells are a result of the dynamic, spatiotemporal interplay of individual cells and between different cell types within its microenvironment. Cells attach to one another and form natural cell-to-cell attachments. The cells and the extracellular matrix that they synthesize and secrete in 3D is made of complex proteins in their native configuration that provide important biological instructions to the cells resulting in more complex phenotypes. Intravital imaging within mice for example show that microenvironment dictates the mode of tumour cell migration; streaming where multiple cells migrate together in streams of single cells (Patsialou et al. 2013) or collectively as a compact group (Friedl et al. 2012). In 3D too, cells form multi-layers which are denser than that of 2D culture and better mimicks how drugs must diffuse through multiple cell layers in the human body before it reaches its intended target. Direct comparison of 2D and 3D spheroid culture screens confirm differences in signaling pathways and response to treatment, (Imamura et al. 2015; Riedl et al. 2017). Therefore the establishment of 3D cultures that best resembles the native tissue for high-throughput screening will greatly improve the accuracy of pre-clinical trials. Toxicology studies have shown 3D spheroids to be nearly on par with *in-vivo* studies for the purposes of testing toxicity of drug compounds (Fey and Wrzesinski 2012). However spheroids are an aggregation of cells with little to no relevant tissue structure. Recently organoids have emerged as a potentially more physiologically and clinically relevant 3D culture model.

An organoid is a miniaturized and simplified version of an organ that is similar in both composition and architecture but can be cultured indefinitely in three dimensions *in-vitro*. Importantly they harbour small populations of genomically stable, self-renewing stem cells that can give rise to fully differentiated progeny comprising all major cell lineages at frequencies similar to those in living tissue.

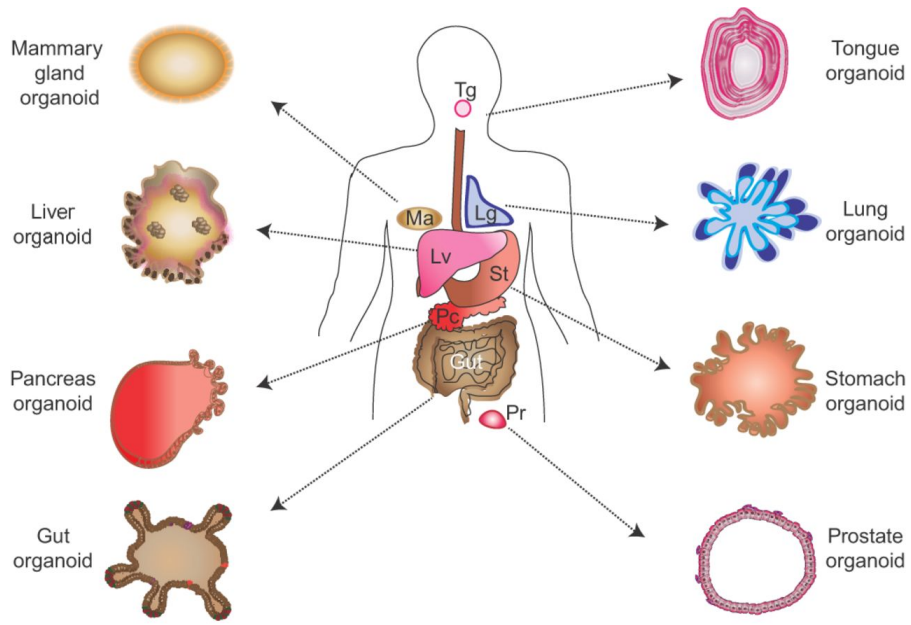

Figure 1.1: The organoid zoo. Plethora of different organoids that have been successfully grown from primary tissue and/or stem cells. Adapted from Huch and Koo 2015.

They can be derived from one or a few cells from tissue biopsy, embryonic stem cells or induced pluripotent stem cells, which can self-organize and differentiate when cultured in 3D in the presence of an extracellular matrix support. A large library of organoids can now be grown from primary tissue, induced pluripotent cells or embryonic stem cells encompassing numerous body organs including brain, kidney, pancreas, oesophagus, stomach and colon (Huch and Koo 2015; Fatehullah et al. 2016). In some cases, 'disease' tissues have also been recapitulated such as diseased oesophageal (Barrett's oesophagus), intestinal and colonic epithelia. More physiologically relevant than monolayer culture models and more amenable to genetic manipulation compared to *in-vivo* models, organoids are an important bridge between traditional 2D cultures and *in-vivo* mouse/human models, (Yin et al. 2016; Bershteyn et al. 2017; Mills et al. 2017). Intestinal organoids have already demonstrated the potential of this approach; analysing stem cell behaviour, modelling pathogen-epithelia interactions, gene editing, disease models and orthotopic transplantations (Sato et al. 2009; Dekkers et al. 2013; Drost et al. 2015; Finkbeiner et al. 2015; Fordham et al. 2013). These successes has led to efforts to create

cryopreserved biobanks of healthy and diseased human organoids for foundational science and translational applications (van de Wetering et al. 2015). At the same time the potential to grow matched normal and diseased organoids from the same patient is of great interest for the development of personalised treatments that can selectively target the diseased tissue with minimal side-effects, (Cantrell and Kuo 2015; Fatehullah et al. 2016; Pauli et al. 2017).

1.1.3 Automated Analysis for High-Content Screening

High-content, high-throughput experimental methods generates a large amount of data that necessitates automated analysis. Presently this is however a major barrier for timelapse datasets. In general there is a lack of simple analytical systems to robustly, sensitively and quantitatively measure both individual and collective cellular motion dynamics in varying experimental conditions for 2D and particularly 3D cultures. A suitable computational method for studying cell population dynamics, including in a medium- or high-throughput manner, should be: (i) robust, i.e. able to handle inevitable variations in image acquisition and experimental protocol; (ii) sensitive, i.e. able to detect motion differences resulting from small changes in environment or stimuli with a minimum number of replicates; (iii) automatic, not requiring manual intervention except from the initial setting of parameters; and (iv) unbiased, able to characterise motion (e.g. as a motion 'signature') with minimal assumptions about motion behaviour. Existing approaches including vertex models (Fletcher et al. 2014; Alt et al. 2017a) (Ch.3.3.3), differential equations (Cai et al. 2007; Markham et al. 2015) (Ch.3.3.3), cellular automata (Mallet and De Pillis 2006; Hatzikirou and Deutsch 2008; Podewitz et al. 2016) (Ch.3.3.3) and cell tracking (Padfield et al. 2011; Meijering et al. 2012; Maška et al. 2014; Schiegg et al. 2015; Masuzzo et al. 2016; Hilsenbeck et al. 2016) (Ch.2.2) have successfully enabled the assessment of specific biological phenomena (such as individual cell motility or stresses between cell-cell contacts) but do not fully meet these four criteria, are limited in scope of application and are difficult to generalise to medium/high content screening. Specifically, vertex models, differential equations and cellular automata

are founded on prior modelling assumptions that may not be satisfied for arbitrary timelapse acquisitions. Cell tracking methods rely on the ability to accurately detect individual cells over time based on image segmentation and are easily affected by such factors as plating density, image contrast variations, marker expression, low resolution image capture or occlusion, especially over long-time acquisitions (e.g. a week). Particle image velocimetry (PIV) (Ch.3.3.3) can extract motion in dense monolayers (Szabo et al. 2006a; Petitjean et al. 2010) but subsequent downstream analysis do not systematically account for variable video quality, collective motion and global phenomena that occur across large groups of cells such as boundary formation.

Existing video analysis of cellular motion are narrow in focus, particularly when considering collective dynamics. Primarily they operate on the independent comparison of single video statistics concentrating on individual cell motion even in dense monolayers. Only a few dynamic parameters of interest are extracted using a small set of videos such as average cell speed and orientation for statistical tests of significance for a limited set of experimental conditions commonly comprising only wild-type and knock-out cells, (Das et al. 2015; Masuzzo et al. 2016). Further few works explicitly attempt to extract a representation suitable as a video descriptor that can exploit the possibility of clustering videos by motion similarity or to detect 'abnormal' motion in the video. Neumann et al. 2006 demonstrate high-throughput RNAi screening by time-lapse fluorescence microscopy of HeLa cells, but their approach of reducing trajectory features by clustering can be non-discriminative, requires prior-knowledge of the classes and the clusters may not hold for new data. Failmezger et al. 2013 relax the need for pre-specified clusters by using a Hidden Markov Model (HMM) to learn the different temporal transition states present in the video. However these HMM methods apply primarily for single cells of which there are many in a single video and the analysis becomes less interpretable if there are no clear biological equivalent with the inferred hidden 'states' in the model. They also do not scale well from single to multiple videos. More realistically

global treatment affects all cell types and certain cell types such as epithelial cells operate under collective migration in confluent monolayers. Cancer cells exhibit both individual and collective modes of migration. There is a lack of works that describe how to represent and measure the complex spectrum of dynamics within a single video. Zaritsky et al. 2012a compute Local-Binary Patterns and velocity based vectors over distinct temporal segments as a spatiotemporal descriptor and showed how this could be used to differentiate between three different experimental conditions. However this study was limited in replicates with a total of 11 videos for treated and untreated conditions and the temporal segments were not automatically found. The descriptors were shown to exhibit clustering of conditions but cannot be further used to describe the collectiveness in motion. With organoids, variability in organoid growth using current culture methods, challenges in fluorescent 3D timelapse imaging due to sensitivity to laser stimulation and difficulties in staining restrict long-time acquisition to label free microscopy (c.f. Ch.6.2). The resulting videos capture a birds-eye view of a 3D process in 2D resulting in a 2.5D image sequence which presents a number of challenges for automated analysis. To the best of my knowledge there does not currently exist reliable commercial, academic or open-source software to segment and track organoids and analyse the branching morphology. Few works attempt any form of video analysis. Instead static snapshots are segmented on demand through manual or automated means for quantification, (Robinson et al. 2015; Walsh et al. 2016; Garvey et al. 2016) or only centroid tracking is attempted using very low resolution videos (Tan et al. 2015). Mathematical modelling is a more established approach to investigate organoid cellular dynamics given a set of experimental hypotheses, (Buske et al. 2011; Buske et al. 2012) but their parameters cannot be fitted to individual video data.

1.1.4 Precision Medicine and Integrative Analysis

The establishment of a high-throughput organoid culture coupled with an automated analysis framework suitable for high-content screening has immense potential for personalised medicine. Existing methods primarily only use viability to screen

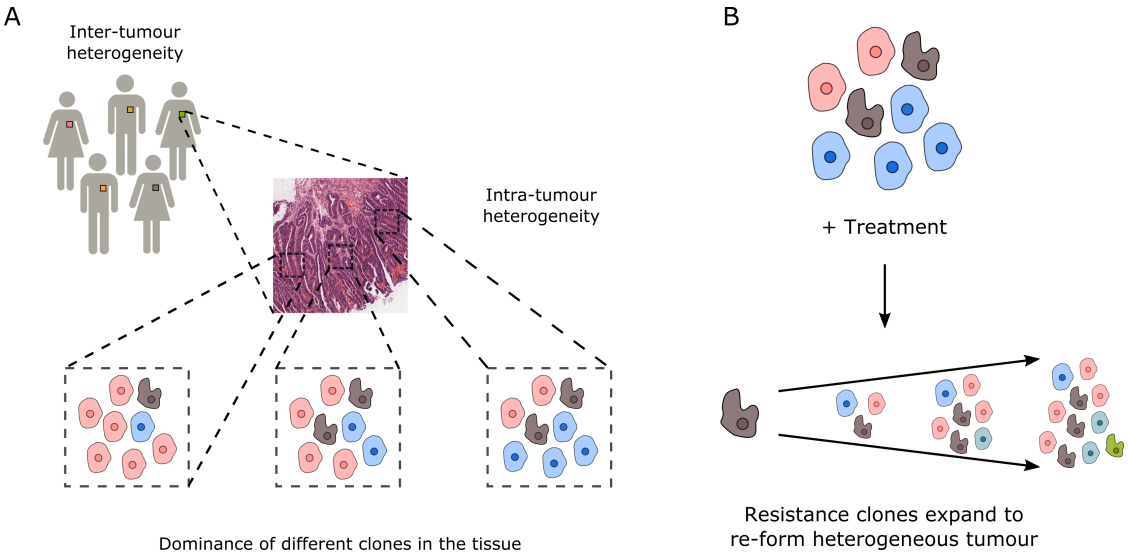

Figure 1.2: The challenge of inter- and intra- tumour heterogeneity for cancer treatment. A) The same tumour can vary between individuals (inter-) and within the same tumour (intra-) with the presence of different clones. B) Following conventional broadband treatment such as chemotherapy or radiotherapy resistant clones expand reforming a heterogeneous tumour.

organoids for further sequencing analysis or use simple shape statistics such as the area to assess growth dynamics. Both of these approaches however neglect the diverse organoid morphologies that are hallmarks of the tissue of origin (Fatehullah et al. 2016), (Fig.1.1) that may be perturbed under pathological conditions. For example van de Wetering et al. 2015 noted that tumor-derived organoids presented with a range of patient-specific morphologies, from thin-walled cystic structures to compact organoids devoid of a lumen. The ability to recommend personalised treatment is particularly important for heterogenous diseases which are a major focus of current research such as Alzheimer's disease, diabetes and cancer. For example cancer is a complex disease involving a progressive sequence of gene-environment interactions that cannot occur without dysfunction in multiple systems, including DNA repair, apoptotic and immune functions. Correspondingly it is highly heterogeneous both within tumours and amongst patients, (Fig.1.2A). In turn this heterogeneity is the primary cause of tumour resistance and the driver of different responses to therapeutic treatment. Current broad spectrum treatments such as chemotherapy and radiotherapy may kill the majority of cells but resistant clones remain which

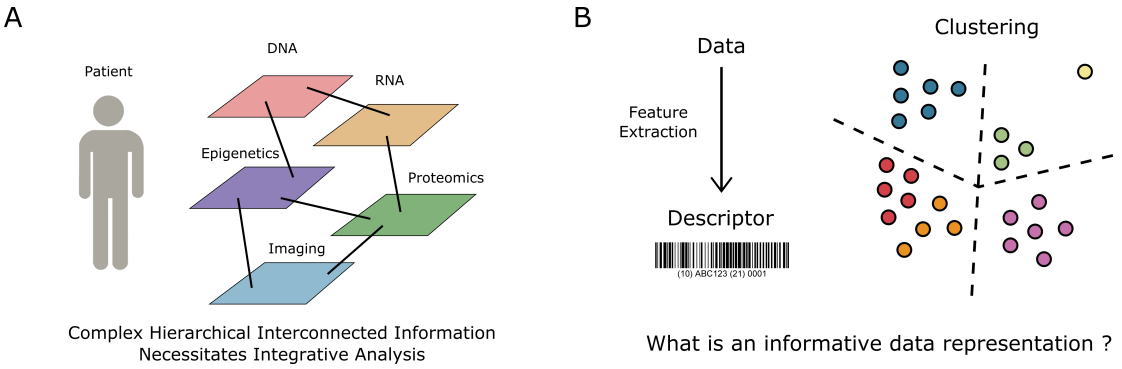

Figure 1.3: Precision medicine requires improved quantitative models for describing individual patient disease. A) Detailed characterisation of the disease status of an individual must account for the complex hierarchy of information from DNA to protein to clinical imaging and understand the interrelation at different levels. B) Data is typically processed and specific features extracted which when concatenated is a descriptor of the data like a barcode for goods. The informativeness of the feature for patient specificity can be assessed through clustering. A good descriptor as a representation of the data clusters together similar patients and can identify outliers.

can expand to reform the heterogeneous tumour causing relapse, (Fig.1.2B).

The understanding of the mechanisms underlying disease heterogeneity is in general extremely complex involving the interplay of multiple factors across different molecular signals (DNA, RNA, proteins including epigenetics and post-translational modifications) and scales (e.g. gene-cell, cell-tissue, tissue-organ, organ-whole person), (Fig.1.3A). In recent years with technological advances particularly in sequencing where now entire normal and cancer genomes can be sequenced in a matter of weeks and at the resolution of single cells, the ability to interrogate disease mechanistically at different molecular levels and to integrate the information hierarchically has emerged. The importance of integration is well recognised (e.g. Barker and Brightling 2013; Ellsworth et al. 2017; Yi et al. 2017). A number of consortia and public databases have arisen that aim to provide a comprehensive collection of molecular information. Some of the best known being the Cancer Genome Atlas (TCGA), the Human Protein Atlas including tissue and cell atlases, the Allen Cell Explorer, the Genotype-Tissue Expression (GTEx) database, the Encyclopedia of DNA Elements (ENCODE), the Catalogue of Somatic Mutations in Cancer (COSMIC) and Genomics England. The construction of integrative,

quantitative descriptors of disease that incorporates information of biological processes at all levels along with temporal dynamics promise to better elucidate the molecular origins and mechanisms of heterogeneity for better risk stratification than can be provided by high-content screening. For videos to be integrated with datasets from other modalities, a feature representation of the video motion is necessary (Fig.1.3B). We hypothesize that an appropriate descriptor for a single video that sufficiently captures the wealth of motion dynamics present such as the collective behaviour between cells, cell morphological changes and interaction between individual cell types is one that facilitates the clustering of videos in a video collection by motion phenotype (Fig.1.3B). As discussed this is a key missing feature of current cell motion analysis algorithms.

1.2 Thesis Aims

We aim to develop a systematic motion analysis framework to handle a diverse set of motion videos with particular focus on cellular motion that uses minimal parameters, is robust to imaging conditions and works out of the box without excessive 'tuning'. It should be suitable for high-content screening of motion. In particular we aim to develop a representation of the entire video motion that not only allows for the assessment of individual cell motion metrics but also accounts for potential emergent collective dynamics and can be used for integration with additional information modalities such as sequencing. The representation should be equipped with a consistent mechanism to quantify local dynamical measures such as average speed within a confluent epithelial sheet and global motion features such as the boundary formation between two epithelial sheets. Specifically, the developed framework should demonstrate the following characteristics:

- **Theoretically motivated:** the basis of analysis and motion extraction are founded on physical principles which provide robustness of performance. End-to-end learning schemes requiring a large dataset with extensive annotation is not pursued in this work.

- **Systematic motion characterization:**

    1. *Individual motion metrics:* such as speed, orientation and other trajectory statistics (Meijering et al. 2012).
    2. *Collective motion metrics:* recognition that cellular movement may be coordinated over a spatial area.
    3. *Local and global phenomena:* systematic method for quantification of localized phenomena such as the movement of cells towards a particular spatial location and global phenomena where cells are coordinated over a large spatial distance such as the formation of sharp boundaries.

- **Motion saliency:** the ability to spatially localise significant motion events in the video, for example spatial locations where cells are attracted to or repelled from such as the fingers of a pianist and the moving boundaries in a wound healing assay.

- **Motion detection and segmentation:** the ability to localise the independently moving objects and demarcate their shape and centroid as they move.

- **Motion classification:** the ability to construct meaningful signatures for each video that allows comparison of the video motion content across multiple experimental conditions and suitable for data integration. We hypothesize signatures that allow good clustering of video collections are also good high-level descriptors of video content and suitable for integration, (Fig.1.3B).

1.3 Overview of Thesis Chapters

An overview of the remaining thesis chapters is given for ease of reading. Chapter 2 reviews general approaches for motion extraction, cell tracking concluding with a general computer vision discussion of ways to represent a video; as a single feature vector, as a sequence and as a collection of moving objects. Chapter 3 reviews motion theoretical frameworks for interpreting and understanding the extracted

motion in a manner that relates to the underlying biological phenomena. Chapter 4 presents our developed framework, motion sensing superpixels (MOSES) that aims to take the first steps towards capitalising upon both the advantages of theory and data-driven machine learning approaches for analysing motion. This is achieved by the local aggregation of motion information using graphs built upon superpixels that partition the image and sample the motion. We demonstrate the core principles of MOSES and its limitations by application to a number of different datasets. Chapter 5 presents an extended case study of MOSES with application to analysing the complex epithelial interactions between heterotypic cell populations in 2D cell culture. We study the interaction dynamics at the human squamocolumnar junction, the interface between the oesophagus and stomach using the respective immortalised *in-vitro* human cell lines. Chapter 6 further develops MOSES for automated analysis of organoid timelapse videos. We show how the framework permits motion segmentation of individual organoid instances in the video using only weak initial spatial positioning to allow the assessment of dynamic morphological changes. We further show how complex organoid branching morphology can be cast to a particle tracking problem and solved within the MOSES framework. Chapter 7 concludes the thesis by summarising the existing advantages and limitations of the system and how MOSES can be extended and improved. In particular we discuss how neural networks could permit end-to-end learning to further scale up MOSES.

1.4 Thesis Contributions

This thesis is the result of a collaborative and interdisciplinary research effort that lies at the interface of biology and engineering. My contributions can be divided into three application domains; biological advancement in the development of novel approaches to analyse a number of biological phenomena, bio-image informatics advancement in the derivation of new metrics to characterise specific biological phenomena and advancements to computer vision in the form of a systematic framework that enables spatiotemporal motion saliency maps, video classification, object segmentation and tracking. The specific contributions and relevant thesis

chapters is summarised in Table.1.1. The work of chapter 5 has been submitted as a journal paper and is under revision. The work of chapter 6 is in preparation for a journal paper. A patent is pending for MOSES.

Biological	**Bio-Image Informatics**	**Computer Vision**
Novel segmentation-free approach for immunosurveillance in zebrafish using spatial motion saliency, (ch.4)	Motion signatures for collective motion, (ch.4,5,6)	A consistent framework, MOSES for i) motion saliency, (ch.4,5,6) ii) object segmentation and tracking, (ch.4,5,6) iii) video classification, (ch.4,5) iv) frame classification, (ch.4) and operates with any modality.
Novel approach for analysing two epithelial cell population interactions, (ch.5)	Boundary formation index, (ch.5)	Casting organoid branching point detection as local strain extrema and application of particle tracking to recover branching (ch.6)
Novel approach for tracking organoid growth and migration including organoid segmentation and branch tracking, (ch.6)	Motion stability index, (ch.5)	
	Mesh disorder index, (ch.5)	

Table 1.1: Table of thesis contributions.

2

Extracting and Describing Motion

Contents

Motion estimation from a sequence of images is the most fundamental step in motion analysis. In this chapter we review the key methods to extract motion from an image sequence for tracking objects. We end by discussing methods for deriving compact video representations that summarise the entire video motion content to enable tasks such as video classification and segmentation. The latter capability is a crucial consideration for data integration as it demonstrates that the derived video signature has the capacity to group similar motion together in the presence of noise such as camera motion, appearance variation and microscope illumination.

2.1 Motion Estimation

There are two predominant methods to estimate motion from a sequence of images i) sparse feature-based methods that track the most distinctive video features and ii) direct estimation methods from spatial-temporal image intensity variations. Feature based methods by tracking only a subset of the image pixels are often faster to run and can better handle changes in viewpoint and large displacements. However being sparse, its performance critically depends on the feature matching performance and is not suited to analyse small motions. Direct motion estimation methods overcome these limitations by estimating the motion for each pixel based on image intensity assumptions between two images. In this thesis we further categorise direct motion estimation methods into integral or differential approaches depending on whether they operate on the summation over or the gradients of image intensities. This is to better highlight the mathematical distinction between particle image velocimetry (PIV), an integral approach that is the most popular method for analysing cellular motion in *in-vitro* cell cultures and optical flow, a differential method widely used in computer vision to extract motion from general image sequences. We review each method in turn with the most emphasis on optical flow, the chosen method for motion extraction in this thesis. The pros and cons of feature tracking and direct methods is summarised in Table.2.1.

2.1.1 Feature Tracking

Given two images, in feature tracking, keypoints in each image are identified, the local image patch is extracted and image features are computed to encode the patch appearance. Then the image features are matched to infer the relative motion between the two images, (Fig.2.1). Inherently this process assumes that the moving object of interest possesses 'interesting' points whose local appearance are preserved temporally and are salient in the image. There is no single definition of an 'interesting' point. Ideally keypoints should be distinctive, and invariant to scale and illumination. In practice corner-like and 'blob' features are used offering spatial localisation and sparsity. They are efficiently located as the extrema of the

Approach	Pros ☺	Cons ☹	Examples
Feature-based	• large displacements • fast computation • robustness to illumination	• small displacements • non-rigid motion	Tomasi and Kanade 1991
Direct motion estimation methods	• small displacements • smooth motion	• slower computation • unsharp boundaries • More illumination sensitive	Farnebäck 2003; Brox et al. 2009

Table 2.1: Pros and cons of feature-based and direct motion estimation.

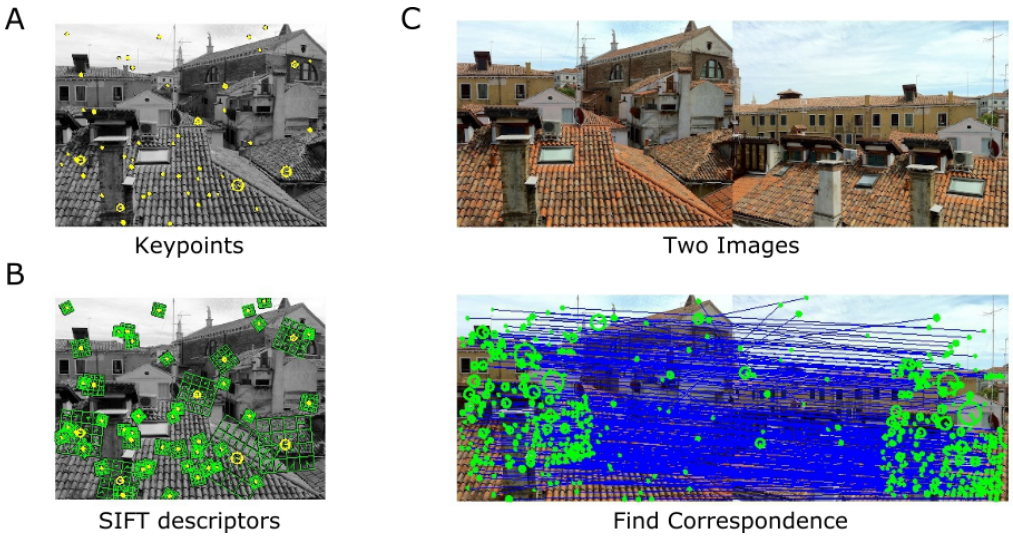

Figure 2.1: Feature based image matching for motion extraction. A) Key interest points are detected in the image. B) Feature descriptors are computed over a local image patch around the detected interest points. C) Extraction of keypoints and corresponding descriptors allows matching across two individual images. Images in A)-C) adapted from VLFeat website, http://www.vlfeat.org/overview/sift.html.

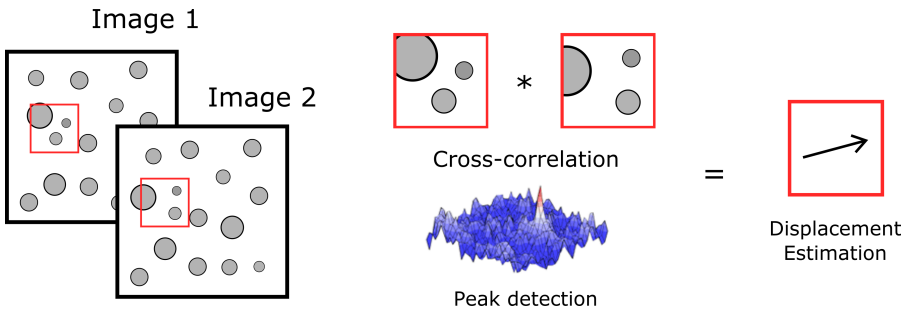

Figure 2.2: The particle image velocimetry pipeline.

2nd image derivative (Hessian matrix). This is the basis of Laplacian of Gaussian (LoG) approaches and the Shi-Tomasi corner detector, (Shi 1994). Perhaps the most well-known feature-based tracker is the seminal Kanade-Lucas tracker (KLT) (Tomasi and Kanade 1991) which even today is widely used as a robust sparse tracker. We describe the basic KLT algorithm. 1) Corners are found satisfying an image Hessian eigenvalue condition as good features to track. 2) For each corner the displacement to the next frame is computed using Lucas-Kanade image registration (Lucas and Kanade 1981). 3) The displacement of each corner is stored and its position is updated accordingly. 4) (optional) Add more corner points every M frames using step 1. 5) Repeat steps 2-3 and optionally 4. 6) The final result is a sparse set of trajectories for each corner point. The basic formulation is not suited for handling large motions. In practice, a multi-scale scheme, the pyramidal KLT tracker (Bouguet 2001) is used.

2.1.2 Particle Image Velocimetry (PIV)

In the fields of fluid and aerodynamics, tracer particles are used to visualise the motion field. PIV developed as a method to obtain accurate global velocity measurements for high-density particle images. For low particle density images, PIV is similar to particle tracking, (Chenouard et al. 2014). For high particle densities, correlation-based PIV is used with no image segmentation involved. In computer vision PIV is also known as block-matching. We describe the latter correlation approach, (Fig.2.2). To find the motion field, image 1 (time t) and image 2 (time $t + \delta t$) separated by a time interval δt is subdivided into small

image patches or interrogation windows. The two-dimensional spatial convolution of a patch from image 1, I_1 and image 2, I_2 respectively with separation vector $\delta s = [\delta x, \delta y]$ is then computed:

$$R(\delta s) = \iint_{Image} I_1(x, y) I_2(x + \delta x, y + \delta y)\, dx\, dy \tag{2.1}$$

This integral can be decomposed into three components, $R = R_C + R_F + R_D$ where $R_C(\delta s)$ is the correlation of the mean image intensities and $R_F(\delta s)$ is the fluctuating noise component. $R_D(\delta s)$ is the displacement-correlation peak, the cross-correlation function that particles in the first image is the identical particles in the second image. This peak is maximum for $\delta s = [\delta x_0, \delta y_0]$ where $(\delta x_0, \delta y_0)$ is the displacement all particles in the interrogation window has moved. The cross-correlation is efficiently computed with the Fast Fourier Transform and sub-pixel accuracy is achieved through peak fitting e.g. assuming a Gaussian profile. Integration is a smoothing operation therefore particle images which exhibit discontinuous illumination variations is well suited to this integral approach. Nevertheless incorrect correlation is highly likely, resulting in spurious vectors. It is common as part of post-processing to remove such outliers (e.g. identifying velocity vectors that differ from neighbouring vectors) to improve the quality of the inferred motion field. Further as the whole interrogation window is assumed to experience the same displacement the accuracy is constrained by the interrogation window size. Differential motion estimation methods bypass these drawbacks and can utilise local and global constraints to systematically correct for spurious vectors.

2.1.3 Optical Flow

Theory and Assumptions

Optical flow is the estimation of motion based on the brightness patterns in an image sequence. Ideally the optical flow should reflect the actual motion however the estimated motion can also be due to lighting changes without any actual motion as illustrated in Fig.2.3A. Nevertheless it works extremely well in practice. Only a concise overview is presented in this chapter. For a more comprehensive recent

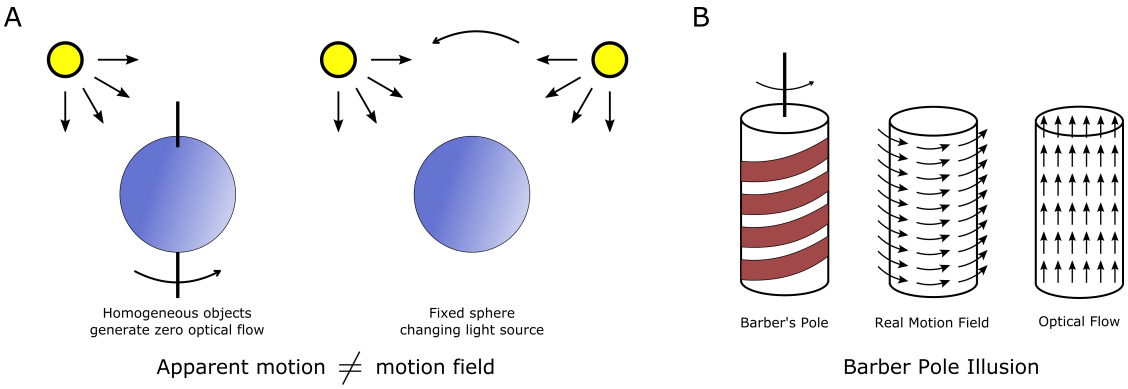

Figure 2.3: Apparent motion is not always the same as the true object motion. A) Left: The image does not change for a smooth sphere rotating under constant illumination but the motion is not zero. Right: A fixed sphere is illuminated by a moving light source. The shading in the image changes giving nonzero optical flow but the true motion field is zero. B) The barber pole illusion and the aperture problem. The bar or contour within a frame of reference provides ambiguous information about its "real" direction of movement (3D rotational motion around axis) and the apparent motion is subject to the shape of the aperture used to view the motion. A vertically elongated aperture accentuates vertical motion, a horizontally elongated aperture accentuates horizontal motion.

review see Fortun et al. 2015.

Given two successive images separated in time by δt with image pixel intensities $I(x, y, t)$ and $I(x+\delta x, y+\delta y, t+\delta t)$ at position (x, y), optical flow estimates for each pixel the velocity vector, $\vec{\mathrm{r}} = (u, v)$ describing the speed and direction the pixel moved across the image. To solve this problem, optical flow makes several key assumptions:

Optical Flow Assumptions

1. Brightness Constancy: the same pixel looks the same in every frame.

$$I(x, y, t) = I(x + \delta x, y + \delta y, t + \delta t) \tag{2.2}$$

2. Small Motion: pixels do not move very far from one frame to the next and the time between successive frames is small i.e. δx, δy and Δt is small, such that using a Taylor expansion:

$$I(x + \delta x, y + \delta y, t + \Delta t) \approx I(x, y, t) + \frac{\partial I}{\partial x}\delta x + \frac{\partial I}{\partial y}\delta y + \frac{\partial I}{\partial t}\delta t \tag{2.3}$$

Subtracting equations Eqn.(2.2) and Eqn.(2.3) yields the fundamental optical flow constraint equation.

$$\frac{\partial I}{\partial x}u + \frac{\partial I}{\partial y}v + \frac{\partial I}{\partial t} = 0, \quad u = \frac{\partial x}{\partial t}, v = \frac{\partial y}{\partial t} \tag{2.4}$$

The gradient terms, $\partial I/\partial x$, $\partial I/\partial y$, $\partial I/\partial t$ are known and can be determined from the data but u and v are unknown. Unfortunately there is only one equation, therefore one cannot solve for u and v uniquely for each pixel. This is known as the *aperture problem* as illustrated by the barber pole illusion in Fig.2.3B where for edges, depending on the viewing aperture the "real" direction of movement is ambiguous. To solve this problem, additional equations need to be generated. This leads to the third assumption, local spatial smoothness.

3. Spatial Smoothness: pixels move like their neighbours. Departure from smoothness E_{smooth} can be quantified by the squared magnitude of the spatial derivatives of the velocity.

$$E_{\text{smooth}} = \iint_{image} \left(\left(\frac{\partial u}{\partial x}\right)^2 + \left(\frac{\partial u}{\partial y}\right)^2 \right) + \left(\left(\frac{\partial v}{\partial x}\right)^2 + \left(\frac{\partial v}{\partial y}\right)^2 \right) dx\, dy \tag{2.5}$$

Together with an error term quantifying the departure from the optical flow constraint, Eqn.(2.4)

$$E_{\text{OptFlow}} = \iint_{image} \left[\frac{\partial I}{\partial x}u + \frac{\partial I}{\partial y}v + \frac{\partial I}{\partial t} \right]^2 dx\, dy \tag{2.6}$$

(u,v) at each pixel position can be determined by minimising the error

$$E = E_{\text{OptFlow}} + \alpha E_{\text{smooth}} \tag{2.7}$$

where α is a constant weighting factor balancing the contribution of the error sources. This is the Horn-Schunck dense optical flow algorithm, (Horn and Schunck 1981).

Optical Flow Algorithms

Briefly we describe three widely available reliable optical flow estimation algorithms in increasing performance that improve upon the basic Horn-Schunck algorithm.

Farnebäck Flow. The Farnebäck flow, (Farnebäck 2003) is a fast, linear algorithm that solves for the optical flow between two images by modelling each image, $I(x, y, t)$ with a polynomial equation, e.g. a quadratic function

$$I(x, y, t) = \begin{bmatrix} x \\ y \end{bmatrix}^T \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} + \begin{bmatrix} b_1 & b_2 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} + c$$

where the matrix coefficients is estimated from a weighted least squares fit to the image intensity values in a pixel neighborhood. By equating the polynomial model of image 2 with a displaced version of image 1 using the constant brightness condition, $I_2(x, y) = I_1(x + \delta x, y + \delta y)$, the authors derive an equation that solves for the displacement at every pixel position $[\delta x, \delta y]$. In practice for robustness to image noise, the authors additionally parametrise the displacement as a polynomial and estimate over local image neighbourhoods and at multiple scales. Due to its local formulation and assumption of a slowly varying motion field, Farnebäck becomes inaccurate for large rigid objects which are uniformly coloured with sharp edges such as a car undergoing large displacements. There is no handling of motion discontinuities. Further, violation of the brightness constancy due to changing illumination will cause all algorithms based only on the constant brightness assumption to perform poorly.

Brox Flow. To improve optical flow estimation in the presence of illumination variations Brox et al. 2009 additionally use the gradient constancy condition Eqn.(2.8) introduced in Uras et al. 1988.

$$\nabla I(x, y, t) = \nabla I(x + \delta x, y + \delta y, t + \delta t) \tag{2.8}$$

Whilst brightness constancy is helpful for translatory motion, gradient constancy is more suited for complex motion patterns. To enable further robustness Brox et al. 2009 refrain from deviations to the linear model, operating instead on the original non-linearised brightness constancy constraint and introduce a multiscale solution where the proposed modified set of optical flow equations, Eqn.(2.9) is solved iteratively first at a coarse image scale obtained from multiple downsamplings

and subsequently propagating the coarser solutions as seed solutions for higher finer image scales until we recover the full solution at the original image resolution. This scheme is known as coarse-to-fine warping. Lastly they use a modified L-1 minimisation, $\Psi(s^2) = \sqrt{s^2 + \varepsilon^2}$ to reduce the influence of outliers on the final estimation.

$$E_{\text{OptFlow}} = \int_{image} \Psi(|I(\mathbf{x} + \mathbf{\Delta x}) - I(\mathbf{x})|^2 + \gamma|\nabla I(\mathbf{x} + \mathbf{\Delta x}) - \nabla I(\mathbf{x})|^2)\,\mathbf{dx} \quad (2.9a)$$

$$E_{\text{smooth}} = \int_{image} \Psi(|\nabla u|^2 + |\nabla v|^2)\,\mathbf{dx} \quad (2.9b)$$

$$E = E_{\text{OptFlow}} + \gamma E_{\text{smooth}}, \gamma > 0 \quad (2.9c)$$

where γ is a constant that weights the two contributions, $|\cdot|$ is the vector magnitude. The smoothness constraint is the same as (2.5) with the modified L-1 minimisation. Minimising E recovers the dense optical flow whose Euler-Lagrange equations is solved efficiently using a two fixed point iteration scheme. The result is a more robust optical flow scheme which can also model large displacements. However it is slower than the method of Farnebäck. For applications that require real-time processing it might be more desirable to estimate the motion faster at the cost of a slight decrease in accuracy.

Dense Inverse Search. To solve optical flow in real-time, dense inverse search was proposed by Kroeger et al. 2016 with very low time-complexity and combines the idea of feature matching methods with the variational solutions of Horn Schunck, (Eqn.(2.7)) and Brox Flow, (Eqn.(2.9)). It consists of three parts. The first image is divided into a regular grid of patches and the displacement for each patch is solved using an optimized inverse search formulation to find the closest looking patch of the same size in the second image. A dense displacement field is then obtained through patch aggregation along multiple scales from the coarse to fine scale as with Brox flow. At each scale, the displacement field is refined using the same objective as Eqn.(2.9). Critically, by operating on the patch-level as opposed to individual pixels combined with an efficient patch correspondence solver, dense inverse search greatly accelerates dense optical flow computation. Of the three discussed optical

flow algorithms, it is the fastest implemented method on a single CPU.

As a fundamental problem, optical flow has been widely researched. The presented algorithms represent only a fraction of those proposed. They are not state-of-the-art but have inspired the design of recent state-of-the-art algorithms. We highlight a few of these advancements. Sift flow (Liu et al. 2011) frames the optical flow solution based on sift features rather than RGB pixel intensities. This is slow for dense correspondences. Large displacement optical flow (LDOF) (Brox and Malik 2011) exploit only sparse feature matching to better guide the Brox variational solution. EpicFlow (Revaud et al. 2015) build upon the sparse matching with edge-preserving interpolation to initialise a dense flow which is refined with variational energy minimization. Consequently sharp boundaries are better preserved. Presently the best performing method on KITTI (Table.2.2) is Instance Scene Flow (ISF) (Behl et al. 2017) based on learning a conditional random field (CRF) model comprising image pixels, smoothness regularisation and instance image segmentations provided by convolutional neural networks (CNNs). On Sintel (Table.2.2) the best performing method is DC-Flow (Xu et al. 2017) which uses CNN extracted image features to construct a cost volume that yields a semi-global matching between two images. The final dense flow is obtained by improving the EpicFlow interpolation scheme using an ultrametric contour map (UCM), an image segmentation hierarchy based on boundary finding (Arbelaez 2006). In summary, good, efficient optical flow is obtained by utilising reliable feature matching, exploiting object boundary cues combined with refinement and densification by local variational optimization. It is interesting to note that the best methods do not employ deep networks for motion estimation. Pretrained image-based CNNs are only used as an optimized feature extractor. However since end-to-end learning is an interesting area of research for completeness we review attempts to utilise deep learning to learn optical flow directly from data.

Dataset	No. Frame Pairs	No. Frames with Ground Truth	Ground Truth Pixel Density	References
Middlebury	72	8	100%	Scharstein and Szeliski 2002
KITTI	194	194	50%	Geiger et al. 2012; Menze and Geiger 2015
Sintel	1,041	1,041	100%	Butler et al. 2012
Flying Chairs	22,872	22,872	100%	Dosovitskiy et al. 2015; Ilg et al. 2016
FlyingThings3D	21818	21818	100%	Mayer et al. 2016

Table 2.2: Table summarising the available optical flow datasets with ground truth frames.

End-to-end Learning Approaches

With recent technological advancements in GPU architectures and the success of deep learning approaches in image classification (Krizhevsky et al. 2012; Simonyan and Zisserman 2014b; Szegedy et al. 2015; He et al. 2016), image segmentation, (Ronneberger et al. 2015a; Badrinarayanan et al. 2015; Zheng et al. 2015), generative modelling e.g. pixelCNN, (van den Oord et al. 2016), generative adversarial neural networks (Goodfellow et al. 2014a; Radford et al. 2015; Odena et al. 2016; Metz et al. 2016) there has been much interest in training neural networks to compute optical flow directly from an image sequence. A key problem of training neural networks to compute optical flow compared to image recognition is the difficulty of obtaining ground truth motion fields for realistic motion. Table.2.2 summarises the popular datasets available to date. One of the first works which reaches similar performance to the classical methods presented above was Flownet, (Dosovitskiy et al. 2015). In the paper, two architectures were proposed, FlowNetSimple (FlowNetS) and FlowNetCorr (FlowNetC). In FlowNetS, a pair of RGB frames is concatenated and fed into a U-net (Ronneberger et al. 2015a) like CNN to predict a downsampled dense optical flow map. In FlowNetC, the two RGB images are fed into two separate streams with tied weights yielding two separate feature maps whose correlation is fed into another CNN to predict a downsampled optical flow map as

in FlowNetS. To overcome data size limitations the authors proposed training on a synthetically generated dataset of chairs which were computationally displaced by a known displacement and a background image added to model realistic cases (Flying Chairs dataset, Table.2.2). The trained networks could subsequently be fine-tuned for a smaller dataset such as Sintel or KITTI. They found whilst the trained networks performed well on synthetic animated movies such as Sintel for which the flying chairs dataset possessed similar motion statistics, in their subsequent work Flownet2.0, (Ilg et al. 2016) both FlownetS and FlownetC could not handle small displacements found in real-world applications such as action recognition. This is somewhat counter-intuitive considering the biggest problem of classical methods was the estimation of large displacements. To overcome this limitation, Flownet2.0 separately trains a network on small displacements called FlowNet-SD which in combination with FlowNetS and FlowNetC greatly improves the optical flow estimation. At present it is the best end-to-end learnt network on the KITTI benchmark dataset. Works such as Deepflow (Weinzaepfel et al. 2013) exploit the hierarchical nature of convolutional networks to extract interest points and deep feature descriptors for patch based matching between two images in order to compute the optical flow but do not learn the filters. Instead of coarse-to-fine as with Brox flow and most multi-scale approaches, this method is sparse-to-dense with sparse matching first implemented at the original image resolution and dense matching at the highest level of pooling.

2.2 Cell Tracking

Biological motion analysis predominantly involve some element of tracking in order to extract cellular motion as cells move singularly or in confluent groups. When cells are fairly sparse, individual cells can be identified and single cell tracking approaches can be applied. Single cell tracking has been successful and has proved essential for understanding the temporal dynamics of cell behaviour e.g. cell lineage tracing and cell shape changes. In particular cell lineage tracing has been a great success for automated tracking for example automatic cell lineage tracing in *Caenorhabditis*

elegans (Bao et al. 2006), in the development of the plant *Arabidopsis thaliana* (Fernandez et al. 2010) and the first cell lineage reconstruction of early *Drosophila melanogaster* nervous system development (Amat et al. 2014). When cells are too confluent the popular literature approach is to use PIV techniques (Sveen 2004; Milde et al. 2012) to extract the dense motion fields for constructing dense trajectories and averaged cellular motion fields. We discuss the analysis of dense motion fields in chapter 3. Here we review developed approaches for single cell tracking which has dominated the field of biological motion research. Single cell tracking is a highly challenging research problem. Each distinct cell must be assigned a unique label and this label must be consistently maintained throughout the temporal motion. Compared to solving this task in the general tracking of humans and objects from everyday videos (Ch.2.3) this may be an even harder problem. Unlike natural everyday activity videos, cells are less distinctive. The same cell type exhibit very similar appearance features with poor contrast to the image background, poorly defined edges and can undergo highly dynamic shape changes. Further there are far more instances of the same object in cell images, 10s-100s of cells compared to 1-10s of humans or objects in an everyday scene.

Single cell tracking approaches build upon the development of general object tracking algorithms which are too diverse to review here (c.f. Smeulders et al. 2014). Broadly, current methods can be classified into deformable models, state-space models and segmentation-based object association. Deformable models utilise a contour evolution approach e.g. a level set or an active contour to obtain the object boundary in the current frame by evolving the contour from the previous frame. The object segmentation is typically initialised in the first frame and updated in subsequent frames. Such contour evolution models fail in applications where objects undergo sufficiently large displacements or show little overlap between consecutive frames, (Lou and Hamprecht 2011). They also fail to detect new cells that enter the field of view. State space models for tracking assume an underlying motion and/or appearance model of the tracked object. The model expects the object to follow an

assumed motion pattern but does not necessarily require an accurate segmentation of the tracked objects. In general they are more computationally demanding due to the large number of hypotheses required to be able to construct sufficiently complex observation models for the motion under scrutiny (Meijering et al. 2009). State space models are, however, better suited to handle larger displacements (Godinez et al. 2011; Ku et al. 2007). Segmentation based object association tracking models involve two major processes: object detection or segmentation in all frames and the subsequent association of objects in different frames to obtain a tracking lineage. This allows cell tracking to be handled as two separate tasks; detection and association. Such association methods have been shown to be effective for cell tracking, to scale well and also achieve high accuracy overall in cell tracking (e.g. Li et al. 2010; Lou and Hamprecht 2011; Padfield et al. 2011; Bise et al. 2011). Today it is perhaps the most popular framework for cell tracking and we discuss some of the developed solutions.

In general there are two common types of imaging modalities for cell tracking; label-free methods such as phase contrast or DIC (differential interference contrast) images which are grayscale and consists of the cells and the imaging background and fluorescently labelled cells where individual cellular components or whole cells are highlighted with dyes of different spectral wavelengths that can be separated and visualized as separate colour channels. In the latter the background is not visible and intensity-based segmentation methods combined with seeded watershed segmentation yields good individual cell segmentations. Recently neural networks have emerged as a potentially more consistent framework for cell segmentation. They show an abilty to extract good features with few training data for individual cell segmentation in both label-free and fluorescence images combined with data augmentation training techniques (Ronneberger et al. 2015b; Van Valen et al. 2016). Importantly the same network can be fine-tuned to enable segmentation of different cell types with few new labelled data. Following segmentation, cells have to be temporally linked to produce the final tracking result. This is arguably the most

difficult process due to the need to account for cells entering or leaving a frame, merges and splits due to cell division and fusion events, considering that cells look very similar compared to objects in natural images. The simplest association method is to link cells to their nearest neighbour across frames. For each frame, the centroid of each cell together with appearance features (image features such as SIFT or statistical morphological features such as area) is extracted. Frame-to-frame each cell is then assigned to its nearest neighbour as determined by their closeness in appearance feature and spatial proximity. This simple approach does not account for biological processes such as cell division and merging. The assignment of cells in successive frames can also be formulated formally as a bipartite matching problem where given two sets we seek to match the elements of one to the other. Jaqaman et al. 2008 solves the linear assignment bipartite problem using the Hungarian algorithm (Kuhn 1955). Padfield et al. 2011 propose a more complex bipartite matching problem, a coupled minimum-cost flow tracking approach where cellular events, *move*, *divide*, *appear* or *disappear* are modelled as additional graph nodes, edge weights are set by image features and an efficient linear programming solution used to solve the minimum flow problem to yield the optimal matching for tracking. Lou and Hamprecht 2011 propose a formal structured learning approach for solving the same minimum flow theory which extends the move set to include *merge* and *split* moves to account for the frequent under-segmentation and over-segmentation errors that occur when segmenting closely-packed objects. To further handle errors due to occlusion, global association methods that link objects over multiple frames by joining multi-frame trajectories is typically used in post-processing. This can be achieved by tracklet stitching (Huang et al. 2008; Huang et al. 2013) where short tracked segments or tracklets are connected via the Hungarian algorithm and dynamic programming (Bonneau et al. 2005). However they do not consider possible cell division and cell segmentation errors. Bise et al. 2011 instead formulate the tracklet association problem as a maximum-a posteriori problem to obtain cell trajectories and lineage trees. The Viterbi algorithm can also be used to achieve a globally consistent set of cell tracks, (Magnusson et al. 2015). Inevitably however any

multi-stage algorithm is suboptimal suffering from irrevocable errors introduced at each sequential step. Recently Schiegg et al. 2015 proposed a single graphical model to jointly model segmentation and tracking. Given two successive frames, cells are first oversegmented using superpixels. Segmentation hypotheses are then constructed based on hierarchical merging over a tree structure. A spatiotemporal graphical model is constructed over all cells and all timesteps with overlapping segmentation hypotheses modelled by intra-frame conflicts and solved. Such a complex model however is difficult to apply and tune for general datasets. Recently neural network approaches for cell tracking have been tentatively explored. Ronneberger et al. 2015a uses U-net and a greedy label propagation algorithm to win the 2015 ISBI cell tracking challenge by a large margin. Critical to the success was the quality of the cell segmentation which took full advantage of the nuances of the dataset in question. In general this strategy is not viable where cell density varies greatly such that cells are no longer well separated, image quality is more variable and cell division and migration are jointly present without either process dominating. He et al. 2017 propose a tracking model where given the cell position in the first frame, a particle filter motion model produces a set of candidate bounding boxes in the following frames. A multi-task learning observation model using CNNs then chooses the best candidate bounding box. An optimized update strategy refines the observation model iteratively to improve tracking performance. Similar approaches that utilise neural networks as a feature extractor after pre-training is the DLT tracker of Wang and Yeung 2013. None of the proposed approaches however can be trained end-to-end. The latter two also only give bounding box or centroid predictions of location with no further shape information.

Single cell tracking remains a major open problem. A major complication is the large biological variability exhibited by single cell motion; some cells migrate more, some divide more, some drastically change shape and the increasingly dominant effect of confluency as cells divide and their boundaries become overlapped. This presents significant challenges for developing a holistic model to capture all the possible

variations across all cell types. The huge lack of well annotated large benchmark datasets for different cell confluencies and cell types comparable to ImageNet (Deng et al. 2009) limits the potential of end-to-end learning approaches. A major focus of this thesis is to investigate situations where solving single cell tracking is not necessarily required for answering the biological question. In Chapter 6 we show a situation where single cell tracking can be effectively utilised to track and monitor individual organoid branches after applying the methods developed in this thesis.

2.3 Video Representations

In its raw form, videos are a chronological sequence of static images. For T frames and image pixel size of $M \times N$ mathematically this is captured by $T \times M \times N$ numbers (dimensions). However this is not a useful way to capture the information content and a highly redundant representation. Even for moderate image sizes the number of pixels becomes very large e.g. $T = 100$, $M = N = 256$ is 6.6 million numbers. Successive frames are highly correlated and the pixels themselves have no notion of the type of motion or the objects within. In video analysis there are numerous cases where we would like to cluster videos by similarity of motion content. Statistically this pixel representation is also ineffective. The large dimensionality suffers from the 'curse of dimensionality' where the few significant differences in pixel values between videos are too insignificant to meaningfully cluster videos without a significantly large number of videos. For these reasons it is useful to mathematically operate on the raw pixels to derive a significantly more compact representation that encodes the useful video information. But this begs the question what aspects of the information should be encoded? and how can we ascertain the usefulness of the representation? For both there is no clear answer. The usefulness of the developed representation is generally assessed through the performance on different analytical tasks such as video classification and object segmentation for which different representations may be more advantageous. We discuss broadly three ways to represent and think of videos in increasing levels of complexity with respect to two computer vision tasks; i) *action recognition*, the classification of a

Dataset	Year	Actions	Background	Clips	Manually Labeled?
KTH	2004	6	Static	600	Yes
Weizmann	2005	9	Static	81	Yes
Hollywood	2008	8	Dynamic	430	Yes
Hollywood2	2009	12	Dynamic	1,787	Yes
Olympic Sports	2010	16	Dynamic	800	Yes
HMDB51	2011	51	Dynamic	6,766	Yes
UCF-101	2012	101	Dynamic	13,320	Yes
ActivityNet-200	2015	200	Dynamic	28,108	Yes
Kinetics	2017	400	Dynamic	306,245	Yes

Table 2.3: Table summarising the common video action recognition datasets.

video collection based on recognition of the portrayed human action and ii) *object recognition* and *segmentation*, the identification and demarcation of independently moving objects in the video spatially and temporally:

Three Ways to Represent Videos

1. Video as a single set of numbers (a single vector): The usefulness is assessed on the action recognition task.

2. Video as a sequence: The usefulness is assessed on the action recognition task.

3. Video as a collection of individual moving objects: The usefulness is assessed based on object recognition and segmentation.

For biological motion analysis most studies only consider the videos as containing a collection of individual moving objects thereby foregoing the potential of the coarse high-level descriptions provided by the feature vector and sequence representations which are more useful for clustering and data integration. Perhaps this is primarily due to the time effort needed to collect a large set of videos for comparison. In the following, we first describe the specific computer vision task followed by the most relevant video representation.

Action Recognition Task

The task of action recognition, the ability to classify a human action based on a short video clip is one of the best studied video classification task in computer vision due to its wide application in sports, security surveillance and search engines. Practically such videos are easy to procure as they are of 'everyday' scenes such as violin playing and easy to associate with as we are very familiar with its content. Computationally three distinct features of this task makes it an interesting problem to study for testing good general representations for video as they are common considerations preserved across all videos. First, a given action can be short, lasting a few frames or long, lasting for multiple frames. The proposed representation should have an appreciation of the temporal visual sequence of events. Second, the action may occur in the presence of background motion and camera artifacts such as shake or zoom. The proposed representation needs to abstract out the salient motion. Third, the same action can be shot and present in different videos from different viewpoints and at different magnifications with different looking humans. The proposed representation must be robust to such appearance variations.

A large number of action recognition datasets have been collected to date with increasing numbers of video clips and actions over time as summarised in Table.2.3. Of the datasets available, the UCF101 and HMDB51 datasets have been the most popular benchmarks at present in the reported literature. UCF-101 (Soomro et al. 2012) is a general video clip collection acquired from YouTube. Each video clip is fairly short and captures one of 101 distinct actions. HMDB51 (Kuehne et al. 2013) was introduced slightly earlier than UCF101 with a smaller set of 51 actions but with clips taken primarily from commercial movies and are therefore less curated and more realistic. Table.2.4 summarises the result of different feature learning methods applied to UCF101 and HMDB51. Handcrafted methods which represent videos as a single vector provide a strong baseline result of 87.9% (UCF101) and 61.1% (HMDB51) but is surpassed by sequential neural network

Method	Approach	Reference	UCF101 Acc. (%)	HMDB51 Acc. (%)
Handcrafted Methods	HoF+FV	(Laptev et al. 2008)	65.9	39.9
	HoG-HoF+FV	(Laptev et al. 2008)	75.4	45.6
	MBH+FV	(Dalal et al. 2006)	81.0	54.7
	GBH+FV	(Shi et al. 2015)	74.2	44.7
	HOG3D+FV	(Klaser et al. 2008)	64.7	38.2
	IDT	(Wang et al. 2013)	85.9	57.2
	IDT + higher FV	(Peng et al. 2016)	87.9	61.1
	IDT + MVSV	(Cai et al. 2014)	83.5	55.9
NN Methods	Deep Networks	(Karpathy et al. 2014)	65.4	-
	Composite LSTM	(Srivastava et al. 2015)	75.8	44.0
	Conv LSTM	(Carreira and Zisserman 2017)	81.0	36.0
	C3D	(Tran-Dinh 2015)	85.2	-
	Two-Stream	(Simonyan and Zisserman 2014a)	88.0	59.4
	TSN	(Wang et al. 2016)	94.2	69.4
	Two-Stream I3D	(Carreira and Zisserman 2017)	**98.0**	**80.7**

Table 2.4: Table summarising the performance of various action recognition algorithms adapted from Caetano et al. 2017. Acronyms are as follows: HoF - histogram of optical flow, HoG - histogram of oriented gradients, MBH - motion boundary histogram, GBH - gradient boundary histogram, FV - Fisher vector, IDT - improved dense trajectories, MVSV - multi-view super vector, NN-neural network, LSTM - long short term memory, C3D - convolution 3D, TSN - temporal segment networks, I3D - inflated 3D.

methods that better exploit the temporal correlations c.f. Two-Stream I3D, 98% (UCF101) and 80.7% (HMDB51).

2.3.1 As a Single Feature Vector

The video is regarded as a single entity and the objective is to summarise the entire video content in a single numerical vector which can be used for classification. Prior to deep neural network approaches, this was the default method for representing videos since the downstream machine learning classifier after feature extraction was typically a support vector machine (SVM) or random forest that require single vector inputs per video. Each video frame is typically treated independently, a series of handcrafted features would be extracted from the image itself and from the corresponding optical flow with respect to the next frame. Subsequently the features

would be concatenated into a single vector. To reduce the resulting vector dimension which can be very large for long videos to better measure similarity, dimensional reduction such as principal components analysis (PCA) or more commonly a feature pooling method such as Bag of Words or Fisher vector encoding would then be applied. We describe the most common set of derived handcrafted motion features below. All (HoG, HoF, MBH and GBH) are based on encoding the image or optical flow appearance using gradient histograms.

Histogram of oriented gradients (HoG). Histogram of oriented gradients was first proposed for images and achieved success in the detection of humans, (Dalal and Triggs 2005). It assumes that local object appearance and shape is well characterised by the image edges similar to SIFT (Lowe 1999). The image is subdivided into small spatial regions ("cells"). Over each cell a weighted 1d histogram is formed based on the magnitude and discretized orientation of the local image gradient. For better invariance to illumination, contrast normalisation of the histogram is applied by accumulating a measure of local histogram "energy". The final descriptor is constructed from the concatenation of the individual cell histograms. For videos a spatiotemporal cell can be used averaged over a few frames.

Histogram of optical flow (HoF). Histogram of optical flow is similar to HoG. Instead of using the image, the optical flow is used instead.

Motion Boundary Histogram (MBH). The optical flow represents the inferred absolute motion between two frames and therefore contains motion from many sources such as foreground object motion and background camera motion. The motion boundary histogram, (Dalal et al. 2006) separates the vertical and horizontal components of the optical flow. HoG-like histograms are derived for each independently and then concatenated into a single feature vector.

Gradient Boundary Histograms (GBH). Instead of image gradients, time-derivatives of image gradients are used instead to emphasize moving edge boundaries. For each frame, image gradients are computed and temporal filtering applied over two consecutive gradient images. The magnitude and orientation of each pixel are then used to build a histogram of orientation as in HOG.

Improved dense trajectories (IDT) recognises the importance of extended temporal continuity and enforces this by computing motion features along tracked trajectories found by tracking a local image patch with dense optical flow. It is the best handcrafted method.

Improved Dense Trajectories (IDT). Dense trajectories, (Wang et al. 2011a) was designed to improve upon the above descriptors by first tracking a dense set of initial points to generate a set of trajectories, $\mathcal{T}$ where each trajectory is the concatenated (x, y) positions of a point (denoted $P_t = (x_t, y_t)$) updated from frame t to $t + 1$ after median filtering M of the dense optical flow field O, $\mathcal{T}_i = \{P_t, P_{t+1}, P_{t+2}, ...\}$ where $P_{t+1} = (x_{t+1}, y_{t+1}) = (x_t, y_t) + (M * O)|_{(\bar{x}_t, \bar{y}_t)}$. Each trajectory is described by a sequence $S = (\Delta P_t, ..., \Delta P_{t+L-1})$ of displacement vectors $\Delta P_t = (P_{t+1} - P_t) = (x_{t+1} - x_t, y_{t+1} - y_t)$. The vector is normalised by the sum of the magnitudes of the displacement vectors $S' = \frac{(\Delta P_t, ..., \Delta P_{t+L-1})}{\sum_{j=t}^{t+L-1} ||\Delta P_j||}$. This vector is known as the trajectory descriptor. The trajectory vector is augmented with HoG, HoF and MBH descriptors computed at each point along the extracted trajectories. In this manner, temporal continuity is naturally enforced. The length of each trajectory is limited to L frames to minimise drift in tracking. After L frames, tracks are removed. New sample points are added to ensure denseness throughout tracking if no tracked point is found within a $W \times W$ pixel neighbourhood. Improved dense trajectories improve upon the original dense trajectories by estimating and removing the camera motion contribution after removal of points belonging to the humans, an idea first proposed in Jain et al. 2016b.

We describe in more detail the two common pooling strategies based on clustering, Bag of Words and Fisher vector encoding that are applied after initial feature extraction for handcrafted methods.

Bag of Words (BoW).

The idea of bag-of-words models was first introduced to capture document word frequencies. Given a finite set of words or vocabulary, the bag-of-words model quantitatively describes each document by the histogram formed by counting the number of occurrences of each word in the vocabulary. For image analysis an image can be regarded as a document from which features around keypoints is detected, extracted and clustered using K-means to define a vocabulary over features. Then as with text documents a histogram can be constructed using the frequency of each feature with nearest neighbour assignment in feature space. Conceptually BoW aims to describe an image with a dictionary of 'prototypes' that is representative of the data. BoW is often applied on the extracted features from the above handcrafted methods to improve performance.

Fisher Vector Encoding (FV).

Instead of K-means clustering for learning the codewords, the Fisher encoding uses GMM (Gaussian Mixture Models) to construct the visual word dictionary yielding for each of the K cluster centres a mean u_k and a covariance v_k vector. The vectors u_k, v_k for each of the K centres is stacked to form the image FV encoding. For very high-dimensional feature vector inputs it is common to first apply PCA to reduce dimensionality before computing the FV encoding. FV in generals delivers better performance than the BoW model.

Due to the use of clustering to learn cluster centres for histogram derivation, both approaches trades off temporal context for greater robustness in the derived features. For example the use of FV yields a 2-3% improved performance over normal IDT, (Table.2.4). Can neural networks with end-to-end learning capability discover a better representation for the data? Deep Networks was one of the first

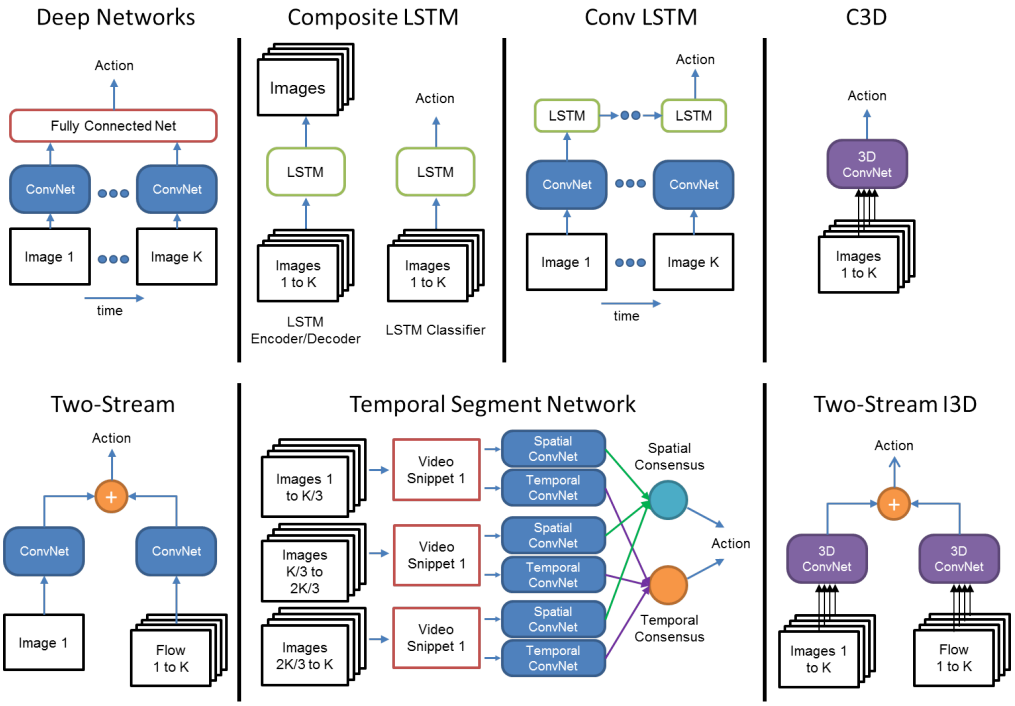

Figure 2.4: Action recognition neural network architectures.

neural network attempts. Karpathy et al. 2014 essentially reformulate the frame-by-frame image feature extraction and feature pooling into a neural network that can be trained with backpropagation. Pretrained CNNs are used to extract a feature descriptor for each of a few sampled video frames and a fully connected layer serves to combine the individual frame features to yield a single feature vector for single action prediction. From Table.2.4, this strategy is only comparable to HoF+FV. This underscores the critical importance of temporal correlations. Subsequent works have aimed to exploit this to various degrees with different architectures. All the different neural network architectures presented is illustrated in Fig.2.4.

2.3.2 As a Sequence

To better exploit the temporal nature of videos, two approaches have emerged that utilise neural network learning. The first idea is the use of recurrent neural network (RNN) modules such as long short-term memory units (LSTM) to sequentially combine and process CNN extracted image features. One of the earliest attempts, composite LSTMs (Srivastava et al. 2015) use a two-step pure LSTM approach

where an LSTM encoder-decoder network is first trained to learn an image feature encoder which is used in a second LSTM network for prediction. LSTMs cannot exploit the hierarchical representation within images that CNNs can and therefore restricts the maximum image area. Applying LSTMs on the extracted CNN features leads to improved but not significantly improved performance on UCF101, Table.2.4 (Conv LSTM). The performance is still not on par with handcrafted IDT. This may be due to information loss in the extracted CNN features. The CNN feature extractor is typically pretrained on static image classification which while yielding strong baseline results, does not fully capture the cues required for temporal recognition. Can we more directly operate on the pixel level information and delay pooling of information until the prediction step? The second idea is to sequentially process spatial-temporal volumes either through the use of stacked inputs e.g. stacking several RGB frames or optical flow images together or via 3d convolutional kernels. Exploiting stacked optical flow and separately processing motion and appearance through two independent CNN networks before fusing their predictions, Feichtenhofer et al. 2016 reported comparable performance to IDT with higher level FV. One might expect based on this that direct optimization over space and time may further improve accuracy. C3D used 3d convolutional kernels directly on the RGB images but do not achieve gains over the two-stream network architecture on UCF101 and HMDB51, (Table.2.4). There are several practical reasons a 3d CNN operating on only RGB images may underperform. One is that the convolutional kernel size dictates the maximum temporal window length to interrogate motion and introduces more parameters to fit compared to image stacking which introduces none. With data overfitting issues, the small size of the UCF101 dataset effectively restrict the network training to smaller temporal kernel sizes and shorter training times. Second, successive RGB input frames looks like jitter. It may be too difficult a problem to expect the neural networks to learn both the motion and appearance of a diverse set of actions starting from random weight initialisations considering the difficulty in learning only the frame-to-frame optical flow (Ch.2.1.3). Temporal segment networks, (Wang et al. 2016) propose a solution

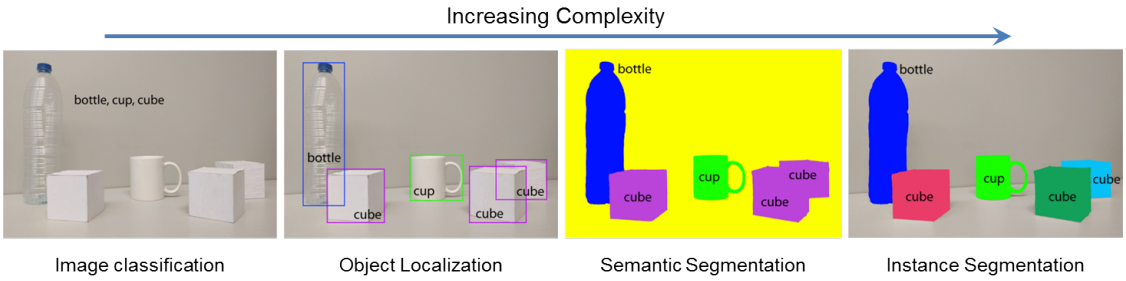

Figure 2.5: Evolution of the complexity of segmentation from coarse-to-fine. Adapted from Garcia-Garcia et al. 2017.

to utilise longer temporal information by breaking the total video length into equal duration temporal segments. From each segment, random video snippets is sampled and fed to two-stream networks to produce snippet-level prediction of action classes. A consensus function such as averaging is then used to separately aggregate spatial and temporal snippet scores into a single video-level action prediction. The authors further explored additional modalities as input to the neural network i) warped optical flow, the optical flow after camera motion correction used in IDT and ii) RGB image differences, a less expensive motion estimation of two frames compared to more expensive optical flow. Presently the best action recognition results reported on UCF101 and HMDB51 is the impressive work of Carreira and Zisserman 2017 whom convincingly demonstrate that the use of a two-stream inflated 3D convolutional network yields significant improvement following pretraining on a larger annotated video dataset, Kinetics. Their work yields two important observations, i) end-to-end learning requires very large annotated datasets, ii) 3d convolutional networks can be built by copying already trained 2d convolutional kernels (inflating) to alleviate overtraining concerns. Wherever possible the provision of additional cues such as optical flow is more advantageous than training on only raw RGB images. In conclusion, the optimal video representation should sufficiently capture both the spatial appearance and temporal continuity in the final video encoding.

Object Detection and Segmentation Task

Object detection and segmentation is the problem of automatically finding the 'important' objects in an image and demarcating their boundaries in an image.

At the same time identified objects should be classified into type e.g. from the view of a moving vehicle automatically identifying and delineating the road, trees, pedestrians etc. For videos this problem needs to be solved from frame-to-frame and the same label given to the same object through time. In general object detection and segmentation is a challenging task that is still intensively being studied. It is often not possible to precisely segment the boundaries of all object instances given the presence of artifacts such as visual occlusion, multiple objects overlapping, illumination changes, objects entering and leaving in addition to background changes. Depending on the precise nature of the question, object detection and segmentation can be achieved at different levels that scales in difficulty, (Fig.2.5): i) detecting the presence of different object classes in the image (classification), ii) detecting every distinct instance of a class using bounding boxes (object localization), iii) delineating the boundaries of each object class (semantic segmentation) and iv) delineating every object instance (instance segmentation). The most well established solution to these problems is to adapt image based techniques where large annotated datasets such as Pascal VOC, (Everingham et al. 2010) and COCO challenges, (Lin et al. 2014) are available. The algorithms are applied frame-to-frame. Label propagation or object association approaches as used in cell tracking above are then used to provide the final consistent labelling across frames. A recent dataset specifically for video object segmentation is the DAVIS challenge (Pont-Tuset et al. 2017). An extensive review of all aspects is beyond the scope of the thesis. We focus on works for object localization and instance segmentation below, the two most relevant for biological cellular motion videos.

2.3.3 As a Collection of Individual Moving Objects

Using object detection and segmentation a video can be viewed as capturing the movement of a collection of distinct objects. Whereas the previous feature vector and sequence representation of videos can be seen as coarse descriptions that is very useful for comparing different videos and for data integration such as with the corresponding audio features, (Jiang et al. 2017), individual object

identification provides the necessary finer granularity to enable the characterisation of motion interactions between individual entities within the video that would not otherwise be captured as distinctly at the global level of video classification. In doing so this representation can help explain why particular videos were clustered together when the feature vector or sequence representation was used. This aspect is particularly important in biological cellular imaging where we wish to probe the dynamic interplay between multiple cell types and for organoid growth where the local number of organoids may lead to competition for growth factors. Given the large variability in biological imaging whilst precise instance segmentation is the most ideal yielding valuable morphological information, this may not be possible in which case the next best alternative is to approximate the localisation via detection. For example, in single cell tracking we might not be able to capture the highly dynamic cell shape changes depending on the staining quality and confluency but we might be able to track the spatial location from the nuclei staining and this information along with an approximate area may be sufficient as input to an improved analysis. In surveillance for example, a facial detector is first applied to approximately locate the areas of individual faces with bounding boxes. The bounding boxes are subsequently cropped, the captured face is aligned to a standard orientation and a facial features algorithm applied to the single faces to classify facial emotion or for facial person recognition. This is termed the Detection-Alignment-Recognition pipeline (DAR) (Huang et al. 2007). Below we discuss object detection followed by instance segmentation. The discussion is most relevant to Chapter 6 for automatic organoid analysis.

Object Detection

The most prevalent approach for object detection is to first automatically generate numerous region proposals where each proposal is likely to contain an object. The bounding boxes of the regions are then computed, image features are extracted and fed into a trained classifier to score the region for object-ness and object classification. Non-maximum suppression is applied in postprocessing to filter out

overlapping bounding boxes. The region sizes are not fixed in the region proposal but are allowed to vary in size in order to allow localisation of all objects of different scales. A key part of the success of this approach is the ability to generate good category-independent region proposals that cover all objects of interest. Prior to deep learning approaches, state-of-the-art methods for generating proposals include objectness (Alexe et al. 2012), selective search (Uijlings et al. 2013), category-independent object proposals (Endres and Hoiem 2010), constrained parametric min-cuts (CPMC) (Carreira and Sminchisescu 2010) and multi-scale combinatorial grouping (Arbeláez et al. 2014). These methods have since been superseded by neural network methods which yield superior performance. The first attempt is that of R-CNN by Girshick et al. 2014 which simply replaces handcrafted feature extraction with that of an ImageNet pretrained CNN network, VGG16 (Simonyan and Zisserman 2014b) and achieves a mean average precision (mAP) of 53.7% on PASCAL VOC 2010. For comparison, Uijlings et al. 2013 achieves 35.1% mAP using the same region proposals, and a handcrafted spatial pyramid and bag-of-visual-words approach. Subsequent work with fast R-CNN (Girshick 2015) and faster R-CNN (Ren et al. 2015) replace the initial R-CNN network with more trainable modules. Notably in faster R-CNN the region proposal could directly be cast as a neural network module to enable the network to learn to generate high-quality region proposals for object detection. Another notable neural network approach is YOLO (You Only Look Once), (Redmon et al. 2016) which is capable of real-time object detection. Similar to faster R-CNN YOLO combines region proposal and classification into one network. Unlike faster R-CNN however, YOLO imposes constraints on the bounding box and models detection as a regression problem. The input image is divided into an $S \times S$ grid. Each grid cell predicts B bounding boxes, confidence for those boxes, and C class probabilities. These are combined to form the final prediction. The resolution of the class probabilities is naturally dependent on the $S \times S$ discretization. The authors report for YOLO more localization errors but reduced prediction of false positives on the background. Object detection is often exploited to guide the instance segmentation solution.

Instance Segmentation

Instance segmentation attempts to identify and classify all the different objects in an image. Its automation is non-trivial, both the number of instances is initially unknown and the evaluation of predictions cannot be conducted in a pixel-wise manner. For static images most progress has been made in the deep learning era. All are based on the idea of first generating regions that localise the objects of interest.

One early solution is the simultaneous detection and segmentation (SDS) method of Hariharan et al. 2014. Bottom-up hierarchical image segmentation and an object candidate generation process, (Multi-scale Combinatorial Grouping) is used to obtain region proposals. For each proposed region, features are extracted using R-CNN fine-tuned on MCG produced boxes and region foreground features and classified by linear SVM followed by non-maximum suppression. Pinheiro et al. 2015 improved upon this with DeepMask, an object proposal method where a CNN predicts jointly for an input patch a segmentation mask and the probability that the patch contains an object. Both tasks are jointly learnt in a single network with the last layers being task-specific. Using this as a starting point the same authors presented an architecture for object instance segmentation via a top-down refinement process, (Pinheiro et al. 2016) that merge low-level features with high-level semantic information from upper network layers to yield better performance in speed and accuracy. Zagoruyko et al. 2016 further developed MultiPath classifier using Fast R-CNN as a starting point and DeepMask object proposals. To improve localization an integral loss was used with foveal regions to provide context and skip connections for multi-scale features. Presently the highest ranked method in COCO segmentation challenge 2016 is that of Li et al. 2016b whom developed a fully convolutional instance segmentation network (FCIS) that can be trained end-to-end. Notably it jointly detects and segments all object instances simultaneously instead of in two separate steps. This is achieved using a score map. Given a region proposal, each pixel within solves two tasks: 1) detection: whether it belongs to

an object bounding box at a relative position (detection+) or not (detection-); 2) segmentation: whether it is inside an object instance's boundary (segmentation+) or not (segmentation-). Importantly their joint formulation fuses the two answers into two scores: inside and outside. By analysing three cases of the inside and outside scores both detection and segmentation is jointly assessed: 1) high inside score and low outside score: detection+, segmentation+; 2) low inside score and high outside score: detection+, segmentation-; 3) both scores are low: detection-, segmentation-. The two scores answer the two questions jointly via softmax and max operations. The region proposals can be supplied by a region proposal network to enable end-to-end training.

For video, a major problem for frame-to-frame instance segmentation approaches is that each image is treated as an independent observation, yet successive frames are highly correlated in information content. For videos with minimal subject occlusion between frames and minimal illumination changes, frame-by-frame segmentation yields acceptable results where the segmentation is generally stable and consistent over multiple frames. However this strategy is unstable for videos where objects constantly leave and enter the field of view. The frame-by-frame segmentation technique no longer find the same salient features for classification. This can be a major problem in biological imaging such as cell tracking, ultrasound and CT where image appearance are more homogeneous and noisy than natural images and conditions can change abruptly between image stacks with uneven illumination, reduced edge contrast and image textural changes. Alternatively temporal information can be explicitly exploited for segmenting moving objects based on the Gestalt principle of common fate: points that move similarly likely belong to the same object. Of particular note is the approach of Brox and Malik 2010 whom extract and cluster the motion trajectories derived from dense optical flow tracking. Like the dense trajectories (Wang et al. 2011a) used for action recognition, seeded points are tracked for as long as possible, provided they satisfy a consistency check; in the absence of occlusion, the backward flow vector points in the inverse direction as

the forward flow vector. New points are uniformly reseeded in the resultant empty areas and tracking is resumed. By virtue of the temporal continuity enforced in the trajectory definition, the resultant segmentation based on spatially regularised spectral clustering of trajectories is also temporally consistent. Occlusions can thus be handled. In addition through temporal continuity, points that initially do not move much can also be grouped. Fragkiadaki et al. 2012 propose the tracing of discontinuities following trajectory embedding for segmentation instead of direct trajectory clustering. Additionally they introduce the notion of context-aware trajectory saliency using the embedded neighbourhood to better separate foreground and background pixels. Finally they densify the trajectory clustering using Gabriel graphs, a subgraph of the Delaunay triangulation based upon an extracted set of line segments using a contour finding algorithm such as Pb (Martin et al. 2004; Maire et al. 2008). The Gabriel graph gives Gabriel superpixels and trajectory clustering induces a labelling on the superpixels where the labelling assignment is solved with graph cuts to give the final instance segmentation. Recently with the introduction of the DAVIS video object segmentation challenge dataset, a few works have explored the extension of image instance segmentation to video instance segmention with neural networks. All works share the basic idea of propagating the initial frame segmentation to subsequent frames. The winning entry in 2017 was Li et al. 2017 whom propose two modules i) mask propagation and ii) re-identification to handle problems of drift, large displacements and occlusion. The modules are trained in a multiple step fashion and then applied to the video sequence in a fairly complex iterative fashion. First an initial segmentation of the objects is obtained from the first frame and these masks are propagated through the entire video. A second pass uses iterative bidirectional (forward and backward in time) application of the re-identification module to separate merged masks and the mask propagation module to re-propagate corrected masks. They report a final global mean (Region Jaccard and Boundary F measure) of 0.699. In comparison the winning entry for the 2016 competition with single object instances in the video scored a final global mean of 0.855. Being the first year of the multiple object challenge, it is safe to assume

there will be more progress to come in terms of architecture and implementation simplification and extensions to end-to-end training.

Having identified and segmented the objects in the video, for biological imaging we would subsequently like to analyse their motion and relate it to potential biological phenomena and physical parameters of interest. In the next chapter we review such quantitative models.

2.4 Summary and Conclusions

- Motion can be extracted from a video by tracking a sparse set of features or agnostically estimated directly from the image intensity variations between successive images using PIV (particle images) or optical flow (general images).
- Single cell tracking is essential for the construction of cell lineage and quantification of shape changes but performance is greatly affected by the quality of image segmentation and temporal linking of cell identity.
- For motion analysis, PIV can be applied to confluent cell monolayers to extract dense motion fields that capture the local motion patterns.
- A video content can be represented:
  1. for classification:
     - as a single feature vector
     - as a sequence
  2. for object level characterization and interaction analysis:
     - as a collection of distinct objects
- State-of-the-art methods for both tasks utilise deep learning approaches that jointly exploit spatial appearance given by the RGB image frames and temporal motion cues provided by dense optical flow.
- For biological motion analysis, cell tracking is an object collection represention. The single feature vector and sequential representation is underexplored in biological motion analysis.
- No consistent framework at present to represent a given video as a single feature vector, as a sequence and as an object collection for video classification and object segmentation.

3

Models for Quantitative Motion Analysis

Contents

The previous chapter discussed methods to extract motion given a sequence of images and how to process the estimated motion to perform video segmentation and object tracking using feature engineering and learning. A key problem with such machine learning methodologies is the difficulty of linking the derived features to the underlying physical and biological phenomena of interest in cellular motion videos. In this chapter we review models derived from mathematical and physical principles that aim to theoretically interpret the extracted motion with biologically relevant measurements. We start by introducing the microscopic theory of transport which

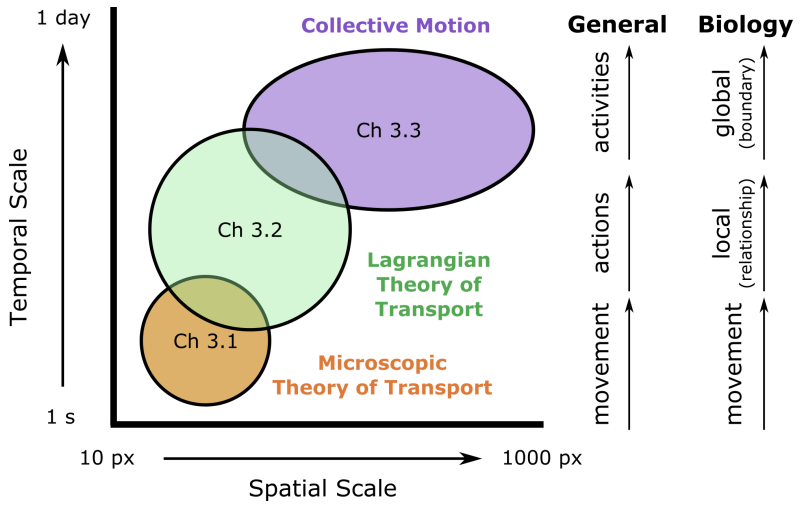

Figure 3.1: Illustration of the spatial and temporal scale of different quantitative motion analysis models. Right of the axis lists examples of the different motion phenomena that can be assessed at the respective temporal scales.

models cells as non-interacting particles. We then discuss Lagrangian theory which analyses how particles move in relation to their local neighbours. Finally we discuss the collective motion phenomena where particles/cells can spontaneously coordinate over large spatial areas and time. These models can be roughly illustrated on a spatio-temporal axis to highlight the spatial and temporal scales of phenomena they typically can be used to capture, (Fig.3.1).

3.1 The Microscopic Theory of Transport

The microscopic theory of movement is based on the consideration of the movement of the particles. It attempts to relate measurable parameters of the global system behaviour such as the diffusion coefficient to the motion of the constituent particles in the fluid (liquid or gas). In particular it theoretically motivates the use of the mean squared displacement (MSD) in the characterisation of diffusion. However it should be noted that using the MSD to characterise single particle trajectories in practice has been found to not always be reliable for motion classification (Briane et al. 2017).

3.1.1 Diffusion and Mean-Squared Displacement

The diffusion of particles is often described with Fick's 2nd law, a phenomenological rule,

$$\frac{\partial n}{\partial t} = -D\nabla^2 n \tag{3.1}$$

where n is the number of particles, D is known as the self-diffusion coefficent and is treated as a material constant. But how does D relate to the atomic motions of the particles within a material or fluid? This problem was first solved by Albert Einstein in 1905 for fluids relating D with the mean squared displacement of a particle.

The mean squared displacement is defined for a system of N particles with positions $\mathbf{r}$ and time t as

$$\langle \Delta^2 \mathbf{r}(t) \rangle = \langle (\mathbf{r}_i(t) - \mathbf{r}_i(0))^2 \rangle \tag{3.2}$$

where $\langle \cdot \rangle$ is the ensemble average and i a particle index. To find an expression for this equation we need to know how the particle positions at the initial time $t = 0$ is related to their positions at time t later. In general we can express this relation in the form of a Taylor series

$$\mathbf{r}(t) = \mathbf{r}(0) + \frac{d\,\mathbf{r}(0)}{dt}t + \frac{1}{2}\frac{d\,\mathbf{r}^2(0)}{dt^2}t^2 + O(t^3) \tag{3.3}$$

Substituting into Eqn.(3.2) we have

$$\langle \Delta^2 \mathbf{r}(t) \rangle = 3v_0^2 t^2 + O(t^3) \tag{3.4}$$

where v_0 is the ideal gas thermal speed, $v_0 = \frac{k_B T}{M}$, k_B is the Boltzmann constant, M is the mass and T the temperature for a system at equilibrium. Thus for short times irregardless of the state of matter the MSD scales quadratically. At long times this behaviour differs in liquids and solids where particles are packed more closely. In a liquid a particle continually undergoes collisions with its neighbours as it diffuses through the liquid whereas in solids the motion is vibratory rather than diffusive as particles are constrained by the lattice forces. We proceed to derive

the Einstein diffusion equation for liquids. In a liquid where no other forces are assumed to act on individual particles other than those due to random collisions, the motion of a liquid molecule is governed by the Langevin equation.

$$m\frac{d^2\,\mathbf{r}}{dt^2} = -\gamma\frac{d\,\mathbf{r}}{dt} + \sigma\zeta(t) \tag{3.5}$$

where m is the particle mass, γ is a viscosity term (Stoke's Law) and $\sigma\zeta(t)$ a noise term with magnitude σ. The corresponding stochastic differential equation given by the Fokker-Planck equations is

$$\frac{\partial}{\partial_t}p(\mathbf{r},t|\,\mathbf{r}_0,t_0) = \nabla^2\frac{\sigma^2}{2\gamma^2}p(\mathbf{r},t|\,\mathbf{r}_0,t_0)$$

If we assume σ and γ are spatially independent we obtain the celebrated *Einstein diffusion equation.*

$$\frac{\partial}{\partial_t}p(\mathbf{r},t|\,\mathbf{r}_0,t_0) = \frac{\sigma^2}{2\gamma^2}\nabla^2 p(\mathbf{r},t|\,\mathbf{r}_0,t_0) \tag{3.6}$$

Noting that the MSD can be expressed in terms of an integral with respect to the particle probability distribution and considering an infinite, volume Ω:

$$\langle(\mathbf{r}(t)-\mathbf{r}(0))^2\rangle = \int_\Omega d^3\,\mathbf{r}(\mathbf{r}(t)-\mathbf{r}(0))^2)p(\mathbf{r},t|\,\mathbf{r}_0,t_0) \tag{3.7}$$

Integrating over Eqn.(3.6) similarly,

$$\frac{d}{dt}\langle(\mathbf{r}(t)-\mathbf{r}(0))^2\rangle = \frac{\sigma^2}{2\gamma^2}\int_\Omega d^3\,\mathbf{r}(\mathbf{r}(t)-\mathbf{r}(0))^2)\nabla^2 p(\mathbf{r},t|\,\mathbf{r}_0,t_0) \tag{3.8}$$

Applying Green's theorem for two functions $u(\mathbf{r})$ and $v(\mathbf{r})$ with an infinite volume and considering the fact that $p(\mathbf{r},t|\,\mathbf{r}_0,t_0)$ must vanish at infinity we have

$$\frac{d}{dt}\langle(\mathbf{r}(t)-\mathbf{r}(0))^2\rangle = \frac{\sigma^2}{2\gamma^2}\int_\Omega d^3\,\mathbf{r}\,p(\mathbf{r},t|\,\mathbf{r}_0,t_0)\nabla^2((\mathbf{r}(t)-\mathbf{r}(0))^2) \tag{3.9}$$

with $\nabla^2((\mathbf{r}(t)-\mathbf{r}(0))^2) = 6$ this is just

$$\frac{d}{dt}\langle(\mathbf{r}(t)-\mathbf{r}(0))^2\rangle = 6\frac{\sigma^2}{2\gamma^2}\int_\Omega d^3\,\mathbf{r}\,p(\mathbf{r},t|\,\mathbf{r}_0,t_0)$$

The right hand side integral evaluates to 1 by definition therefore

$$\langle\Delta^2\,\mathbf{r}(t)\rangle = 6Dt \tag{3.10}$$

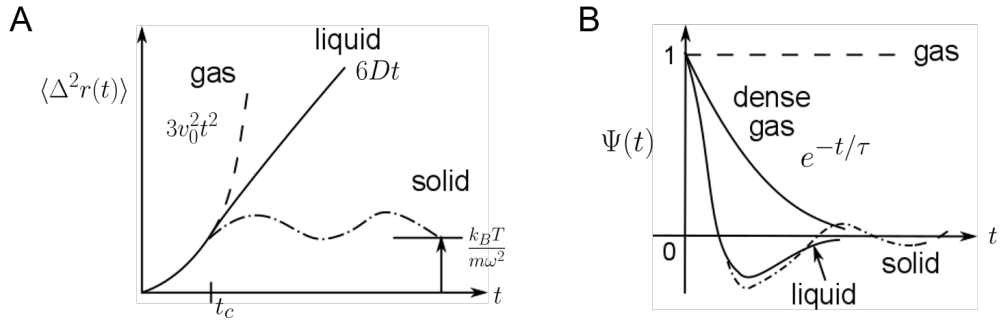

Figure 3.2: Ideal curves for solids, liquids and gases. A) Mean squared displacement curves. B) Velocity autocorrelation curves.

where $D = \frac{\sigma^2}{2\gamma^2}$ is the diffusion coefficient. Thus for a liquid, at long times D varies linearly with MSD and can be directly estimated from the particle positions.

$$D = \left[\frac{1}{6t}\langle\Delta^2\,\mathbf{r}(t)\rangle\right]_{t\to\infty} \tag{3.11}$$

The characteristic shape of the MSD for the three idealized systems of gas, liquid and solid is illustrated in Fig.3.2. The primary feature is a quadratic time dependence at all times for the gas, for the liquid a linear time relationship at long times, and for the solid a relaxation to $k_BT/m\omega^2$ where ω is the vibrational frequency of the solid.

Using the mean squared displacement to characterise the diffusive behaviour of cells is a popular method where the MSD can be described by a power law with respect to time, $\langle\Delta^2\,\mathbf{r}(t)\rangle \propto \Delta t^\alpha$ and the exponent α is used to indicate subdiffusive (α<1), diffusive (α=1), superdiffusive (α>1) and 'ballistic' (α=2) motion (c.f. Ch.5.2.2). However there are several drawbacks to this approach in practice that limits its applicability. Strictly speaking, MSD should be applied to singularly tracked particles which can be difficult given the available imaging quality. Secondly, cells are not point-like and have area which can diffuse. Thirdly, cells can undergo events such as collisions and fusions with other cells. This necessitates other approaches to estimate diffusion coefficients. One approach is through exploiting correlation functions as discussed below. More recently Basset et al. 2017 demonstrated how to accurately recover biophysical parameters such as

the diffusion coefficient from noisy fluorescent images and in the presence of cellular events by appropriate modification of Fick's 2nd law (Eqn.(3.1)).

3.1.2 Diffusion and Time Correlation Functions

The diffusion coefficient can also be linked to the time correlation between individual particle positions (Green 1954; Kubo 1957). Consider the mean square displacement of a typical particle in a fluid at equilibrium, $\langle \Delta^2 \mathbf{r}(t) \rangle = \langle [\mathbf{r}(t) - \mathbf{r}(0)]^2 \rangle$. Using the relation, $\mathbf{r}(t) - \mathbf{r}(0) = \int_0^t \mathbf{v}(t')dt'$ we have,

$$\langle \Delta^2 \mathbf{r}(t) \rangle = \int_0^t dt' \int_0^t dt'' \langle \mathbf{v}(t') \cdot \mathbf{v}(t'') \rangle \tag{3.12}$$

The term in the integrand is known as the velocity autocorrelation function. It can be shown after using the change of variable, $\tau = t' - t''$, $d\tau = -dt''$ and interchanging the order of integration that

$$\langle \Delta^2 \mathbf{r}(t) \rangle = 6v_0^2 \int_0^t dt'(t - t')\psi(t') \tag{3.13}$$

where $\psi(t) = \frac{\mathbf{v}(t)\mathbf{v}(0)}{\mathbf{v}(0)\mathbf{v}(0)}$ is the normalized velocity autocorrelation function, $\langle \mathbf{v}(0)\mathbf{v}(0) \rangle = 3v_0^2$ and $v_0 = \sqrt{\frac{k_B T}{m}}$ the thermal speed of an ideal gas. From Eqn.(3.11) we know that the self-diffusion coefficient D is related to the MSD therefore

$$D = v_0^2 \int_0^\infty \psi(t)dt \tag{3.14}$$

The area underneath the normalized velocity autocorrelation is the diffusion coefficient. The significance of this equation is in relating the thermal particle fluctuations in the fluid at equilibrium to an out of equilibrium global transport coefficient. Similar relationships hold for other transport coefficients such as viscosity and are all mathematically captured by the Green-Kubo relations, (Green 1954; Kubo 1957).

The shape of the autocorrelation function reveals the degree of interaction of particles within an isotropic system with different characteristic shapes for different states of matter, (Fig.3.2). For a dense gas there are many collisions

and $\psi(t)$ exhibits a characteristic exponential decay. As the system becomes more dense, $\psi(t)$ decreases faster in time due to more frequent collisions. For a liquid, molecules are in constant interaction with nearby neighbours and the forces exerted by these neighbours are sufficiently strong to keep the molecule 'caged' within a local region. Large fluctuations in the liquid prevent long term capture and the trapped molecule is able to diffuse away. The 'cage effect' is revealed by $\psi(t)$ becoming negative over a certain time interval.

3.2 Lagrangian Theory of Motion

The mean squared displacement and autocorrelation functions from the microscopic theory of transport is very popular and widely used to characterise the bulk motion properties of a fluid and cell monolayer dynamics with a single number such as the diffusion coefficient or the autocorrelation decay constant. However these measures lack spatial information about the motion and cannot readily reveal irregularities in the motion. The Lagrangian theory of motion provides a localised spatial-temporal perspective of motion by considering how particles move with respect to their local neighbours.

In fluid dynamics there are two primary ways of specifying the fluid motion field, the Eulerian and the Lagrangian specifications. The Eulerian specification looks at fluid motion focusing on specific locations in the space through which the fluid flows as time passes, whilst the Lagrangian specification follows individual fluid parcels through time as they move. The position of an individual parcel through time gives the pathline of the parcel. With analogy to fluid flow, video motion can be quantified through the Lagrangian perspective by following a set of seeded 'particles' whose motion is driven by local optical flow to produce a set of particle trajectories, (Ali and Shah 2007; Wu et al. 2010).

Given a video clip represented as a $T \times W \times H$ where T is the number of frames and $W \times H$ is the frame resolution in pixels (width by height). Let the

corresponding optical flow be (u_i^t, v_j^t) where $i \in [1, W]$, $j \in [1, H]$ and $t \in [1, T-1]$. The position vector (x_i^t, y_j^t) of the particle at grid point (i, j) at time t is estimated by solving the following set of dynamic equations:

$$\frac{dx_i^t}{dt} = u_i^t \tag{3.15}$$

$$\frac{dy_j^t}{dt} = v_i^t \tag{3.16}$$

This can be solved numerically using first order methods such as Euler's method or any other suitable numerical integration method such as Runge-Kutta:

$$x_i^{t+1} = x_i^t + u_i^t \Delta t$$

$$y_j^{t+1} = y_j^t + v_i^t \Delta t$$

where $\Delta t = 1$ is the time separation between frames. Iteratively solving the equation for the seeded particles in each frame, the video's particle trajectory set, $\mathcal{T}$ is obtained summarising all the motion that occur in the video.

$$\mathcal{T} := \{(x_i^t, y_j^t) | i \in [1, W], j \in [1, H], t \in [1, T]\} \tag{3.17}$$

The solution of Eqn.(3.16) can also be viewed in the theory of dynamical systems as a *flow map*, $\phi_{t_0}^t$ which maps a point from their position (x^{t_0}, y^{t_0}) at time t_0 to their position at time t.

$$\phi_{t_0}^t : (x^{t_0}, y^{t_0}) \mapsto \phi_{t_0}^t(x^{t_0}, y^{t_0}) = (x^t, y^t) \tag{3.18}$$

Such a flow map satisfies the following properties.

$$\phi_{t_0}^{t_0}(x, y) = (x, y), \tag{3.19}$$

$$\phi_{t_0}^{t+s}(x, y) = \phi_s^{t+s}(\phi_{t_0}^s(x, y)) = \phi_t^{t+s}(\phi_{t_0}^t(x, y)) \tag{3.20}$$

3.2.1 Finite Time Lyapunov Exponent (FTLE) Field

The Lyapunov exponent is an asymptotic quantity which measures the extent to which infinitely close particles separate in an infinite amount of time (Shadden et al. 2005). It is well defined for periodic fluid flow. For aperiodic flow as is the case for

videos, the finite-time Lyapunov exponent is used to approximate the mixing and dispersion of particles (Shadden et al. 2005). Finite time exponents depends on the initial positions of the trajectories and the length of time over which the particles are tracked, also referred to as the integration time ΔT. A finite time Lyapunov Exponent (FTLE) field is the result when the finite time Lyapunov component is computed over a grid of particles. Let us denote $\mathbf{r}(t) = (x^t, y^t)$. Consider the advection of a particle at $\mathbf{r}$ by the flowmap, $\phi_{t_0}^t$ with an initially small perturbation, $\mathbf{r}' = \mathbf{r} + \delta\mathbf{r}(0)$ where $\delta\mathbf{r}(0)$ is infinitesimal and arbitrarily oriented. After a time interval ΔT this perturbation becomes (Shadden et al. 2005):

$$\begin{aligned}\delta\mathbf{r}(\Delta T) &= \phi_{t_0}^{t+\Delta T}(\mathbf{r}') - \phi_{t_0}^{t+\Delta T}(\mathbf{r}) \\ &= \frac{d\phi_{t_0}^{t+\Delta T}(\mathbf{r})}{d\mathbf{r}}\delta\mathbf{r}(0) + O(||\mathbf{r}(0)||^2)\end{aligned} \tag{3.21}$$

Linearizing the perturbation by dropping higher order terms, the magnitude of the perturbation is given by the standard Euclidean norm as

$$||\delta\mathbf{r}(T)|| = \sqrt{\left\langle \delta\mathbf{r}(0)\frac{d\phi_{t_0}^{t+\Delta T}(\mathbf{r})}{d\mathbf{r}}^* \frac{d\phi_{t_0}^{t+\Delta T}(\mathbf{r})}{d\mathbf{r}}\delta\mathbf{r}(0)\right\rangle} \tag{3.22}$$

where $*$ is the matrix transpose operator and

$$\triangle = \frac{d\phi_{t_0}^{t+\Delta T}(\mathbf{r})}{d\mathbf{r}}^* \frac{d\phi_{t_0}^{t+\Delta T}(\mathbf{r})}{d\mathbf{r}} \tag{3.23}$$

$\triangle$ is known as a finite-time version of the (right) Cauchy–Green deformation tensor. Maximum stretching between particles occurs when $\delta\mathbf{r}(0)$ is aligned with the eigenvector associated with the maximum eigenvalue of $\triangle$, denoted $\lambda_{\max}(\triangle)$.

$$\max_{\delta\mathbf{r}(0)} ||\delta\mathbf{r}(T)|| = \sqrt{\lambda_{\max}(\triangle)}||\bar{\delta\mathbf{r}}(0)|| \tag{3.24}$$

where $\bar{\delta\mathbf{r}}(0)$ is the aligned perturbation. The above can also be rewritten,

$$\max_{\delta\mathbf{r}(0)} ||\delta\mathbf{r}(T)|| = e^{\sigma_{t_0}^{\Delta T}(\mathbf{r})|\Delta T|}||\bar{\delta\mathbf{r}}(0)|| \tag{3.25}$$

where

$$\sigma_{t_0}^{\Delta T} = \frac{1}{|\Delta T|}\ln\left(\sqrt{\lambda_{\max}(\triangle)}\right) \tag{3.26}$$

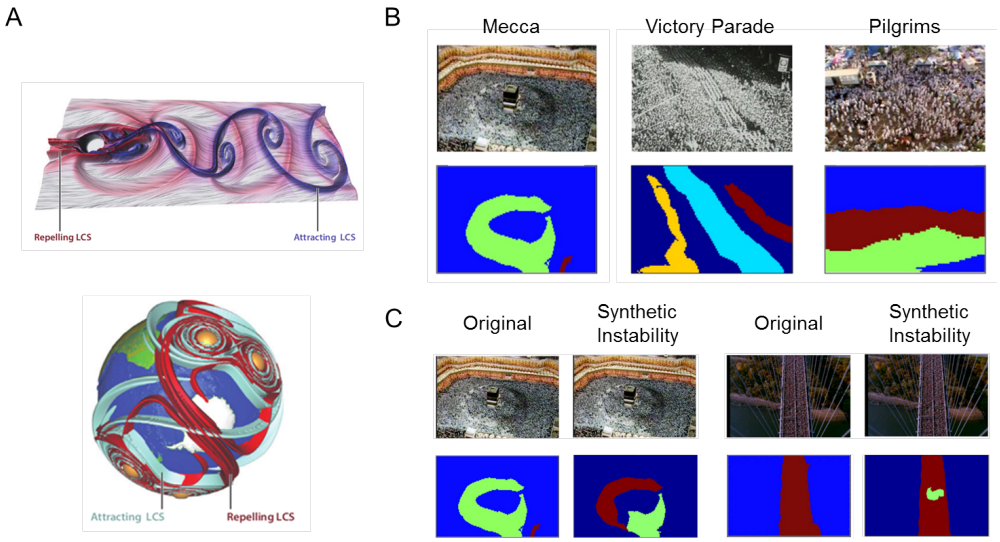

Figure 3.3: Finite time Lyapunov Exponent fields (FTLE), Lagrangian coherent structures (LCS) and motion segmentation. A) Examples of forwards FTLE (repelling LCS) and backward FTLE (attracting LCS). Top: von Kármán vortex sheet behind a cylinder, from Kasten et al. 2010. Bottom: A perturbed four-vortex-ring model for the 2002 splitting of the Antarctic ozone hole, from Lekien and Ross 2010. Ridges of the forward and backwards FTLE field define the LCS. B) FTLE and LCS can be used for motion segmentation and C) unusual motion detection. B), C) adapted from Ali and Shah 2007.

Eqn.(3.26) is the largest finite-time Lyapunov exponent for a finite integration time ΔT corresponding to a particle at time t_0. Note that the absolute value of the integration time is used, permitting both forward and backwards integration times. Computing Eqn.(3.26) with the forward tracks yields the forward FTLE field, with the backward tracks it yields the backward FTLE field.

3.2.2 Lagrangian Coherent Structures

The finite time Lyapunov exponent reveals consistent underlying flow structures that are typically not evident from static snapshots of the motion field. Lagrangian coherent structures (LCS) are defined as the ridges of the FTLE field (Haller 2002; Shadden et al. 2005). The forward FTLE computed from the forward propagation of particles in time reveals repelling Lagrangian coherent structures (stable manifolds for time-independent vector fields) and the backward FTLE form backwards propagation of particles in time locates attracting Lagrangian coherent structures (unstable manifolds for time-independent vector fields), Fig.3.3A.

Essentially they locate the 'sinks' and 'sources' of motion respectively and separate the motion into distinct regions of flow in analogy to edges in images. This property has been exploited successively to segment crowd motion, Fig.3.3B and detect anomalous movements in crowds, Fig.3.3C (Ali and Shah 2007). Vortex-like dynamics can also be investigated using LCS, (Shadden et al. 2006; Green et al. 2010). As a dynamic property of the system, the resolution of LCS in practice depends on the length of the integration time, ΔT (Shadden et al. 2005). If $T \to 0$, the FTLE is instantaneous. For finite ΔT, we obtain a Lagrangian measure of separation as the FTLE considers the integrated effect of the flow over the time interval, (Shadden et al. 2005). However, the ridges in the FTLE field can become more or less pronounced by definition. Some LCSs exist only over short finite-time intervals and will be undetected or detected less robustly if the integration time, ΔT is greater than the event duration.

3.2.3 Chaotic Invariants Features

LCS reveal local structure in the video motion through long-time particle advection suitable for segmenting regions of consistent motion but it is unclear how to construct a set of features from LCS to classify the entire video. Ali et al. 2007; Wu et al. 2010 motivated by Chaos theory, the branch of mathematics which studies how the dynamics of a system changes with respect to the initial conditions propose the construction of chaotic invariants features from the Lagrangian particle trajectories to characterise the video motion.

The Lagrangian particle trajectories form a dynamical system which can be represented in the form of state space models, where the state variables $X(t) = [x_1(t), x_2(t), ..., x_n(t)] \in \mathbb{R}^n$ defines the status of the system at a given time t with n particles. The space, $\mathbb{R}^n$ comprising the state variables is called the *phase space*. The collection of all trajectories from all possible starting points in the phase space of the dynamical system is called a *phase portrait*. An *attractor* is defined as the region of the phase space to which all the trajectories converge to as time approaches

infinity, $t \to \infty$. An attractor is termed *strange* if it is not stable. The invariants of a system's attractor are measures that quantify the properties that are invariant under smooth transformations of the phase space or control parameters. Invariants are grouped into three classes: 1) metric 2) dynamical and 3) topological. Metric invariants include dimensions of different kind and multi-fractal scaling functions, while dynamical invariants include Lyapunov exponents. Topological invariants generally depend on the periodic trajectories that exist in the strange attractor, (Ali et al. 2007).

Taken Delayed Embedding Theorem

In Chaos theory, embedding is defined as a process of mapping a T-dimensional signal where T is the total time to a m-dimensional signal. It allows for the study of systems for which the state space variables and the governing differential equations are unknown. The key underlying observation is that all the variables of a dynamical system influences one another, such that every subsequent point of the given time series results from an intricate combination of the influences of all the true state variables of the system. This observation allows one to introduce a series of substitute variables to obtain the whole m-dimensional phase space, where substitute variables carry the same information as the original variables of the system, (Ali et al. 2007). In practice the embedding is achieved using Taken's delay embedding theorem (Takens 1981) which proposes to construct a new time series m long as the substitute variables from the empirical observations by sampling the original time series at a constant delay time, τ.

$$z_t, z_{t+\tau}, z_{t+2\tau}, ..., z_{t+(m-1)\tau} \tag{3.27}$$

For optimal m and τ these delay vectors generates a phase space for a single trajectory that has exactly the same properties as the original/true variables of the system. Mutual information and false nearest neighbour algorithms can be used to determine the optimal m and τ values, (Perc 2005; Ali et al. 2007; Wu et al.

2010). In Wu et al. 2010 iterative clustering is then applied to find a representative set of trajectories. The embedded sequence is described by computing chaotic invariants to describe the video motion. In Wu et al. 2010 the embedded trajectory is described by the three measures, $F = \{L, D, M\}$ where L is the largest Lyapunov exponent, D the correlation dimension and M is the mean (x, y) coordinate for a representative trajectory. The dimension gives an estimate of the complexity of the system whereas the Lyapunov exponent gives an estimate of the level of chaos in the dynamical system, (Rosenstein et al. 1993).

A representative trajectory with (x, y) coordinates is treated as two scalar time series, one in x, one in y. Chaotic invariants are computed separated for each time series and concatenated. Taking the x-coordinate as example, reconstruct the corresponding trajectory in phase space given by the Taken embedding

$$\tilde{X}_t = [X_t, X_{t+\tau}, X_{t+2\tau}, ..., X_{t+(m-1)\tau}], \quad \forall t \in [1, T] \tag{3.28}$$

Then for a reference point $\tilde{X}_j$ locate its nearest neighbour $\tilde{X}_{\hat{j}}$ by searching for the point that minimises the Euclidean distance, $\hat{j} = \arg\min ||\tilde{X}_j - \tilde{X}_{\hat{j}}||$. Consider in this way each pair of neighbours as nearby initial conditions for the different trajectories.

Largest Lyapunov Exponent

The largest Lyapunov exponent provides quantitative information about trajectories that start initially close together but diverge over time. The video motion is assumed to be chaotic i.e. $L > 0$. To ensure this, Wu et al. 2010 propose using the method of Rosenstein et al. 1993 to estimate L from small and noisy data sets. From the definition of Lyapunov exponents (Eqn.(3.25)) the j-th pair of neighbours can be assumed to diverge approximately at a rate L,

$$d_j(t_i) \approx k_j e^{Lt_i} \tag{3.29}$$

where $t_i = i\Delta t$, Δt is the sampling period of the time series, k_j is the initial separation and $d_j(t_i)$ is the distance between the j-th pair of nearest neighbours after i discrete

time steps. This equation can be rewritten in the form of a set of parallel lines each with slope approximately proportional to L which we wish to estimate.

$$\ln(d_j(t_i)) \approx \ln(k_j) + Lt_i \tag{3.30}$$

The largest Lyapunov exponent is then approximated by fitting the average line defined by

$$y(t_i) = \frac{\langle \ln d_j(t_i) \rangle}{\Delta t} \tag{3.31}$$

where $\langle \cdot \rangle$ denotes the average over all j lines.

Correlation Dimension

The correlation dimension measures the size of an attractor, which defines the chaotic dynamics, and can be estimated from the correlation sum

$$C(\delta) = \frac{2}{Q(Q-1)} \sum_{i \neq j} H(\delta - ||\tilde{X}_i - \tilde{X}_j||) \tag{3.32}$$

where H is the Heaviside step function, δ is a threshold distance and Q is the number of points in the time series. The correlation dimension D is approximated as

$$C(\delta) \approx \delta^D \tag{3.33}$$

3.3 Collective Motion Analysis

The majority of cells in the human body are epithelial in origin. The skin is the largest organ in the body and ∼80% of all cancers originate from the epithelium, (Alberts 2017). Epithelial cells do not usually operate in isolation but are packed tightly in a sheet and move collectively. These processes are fundamental not only for initial tissue development but also the homeostatic maintenance once formed. More broadly, collective motion arises naturally in many physical and biological systems such as the flocking of birds, swimming of schools of fish, migrating bacteria, moulds and ants and has been intensely studied by many different scientific communities (Vicsek and Zafeiris 2012). Interestingly collective motion appears

to be a remarkably universal self-organisation phenomenon. Many studies have shown that certain patterns can arise from the interaction amongst a large number of similar units irrespective of the complexity of the individual units bridging many orders of magnitude. Extremely complex units e.g. cells, cars and people can produce relatively simple patterns of collective behaviour, much simpler than the structure of the single unit itself. Generally collectively moving entities may exhibit only a few characteristic motion patterns (Méhes and Vicsek 2014).

1. Disordered motion. The direction of the motion of individual units is not correlated.

2. Ordered motion. The direction of the motion of individual units is correlated.

3. "Turbulent" motion. There is local order but is lost on a scale much larger than the size of the units.

4. "Streams" of motion. Different streams flow opposite to each other.

5. "Jamming" motion. The restricted area/volume and mutual "pushing" of the units results in a highly strained, locally fluctuating but globally not moving groups of particles.

The Lagrangian theory of motion accentuates local structure to reveal global patterns in the flow but do not explicitly quantify the global motion patterns such as how collective the particle motion is in a group. Here we review some of the theory in terms of measuring collective motion, the biological implications of collective behaviour and discuss how the measurements relate back to the underlying biology. For a more extensive review and introduction we defer to the writeup of Vicsek and Zafeiris 2012.

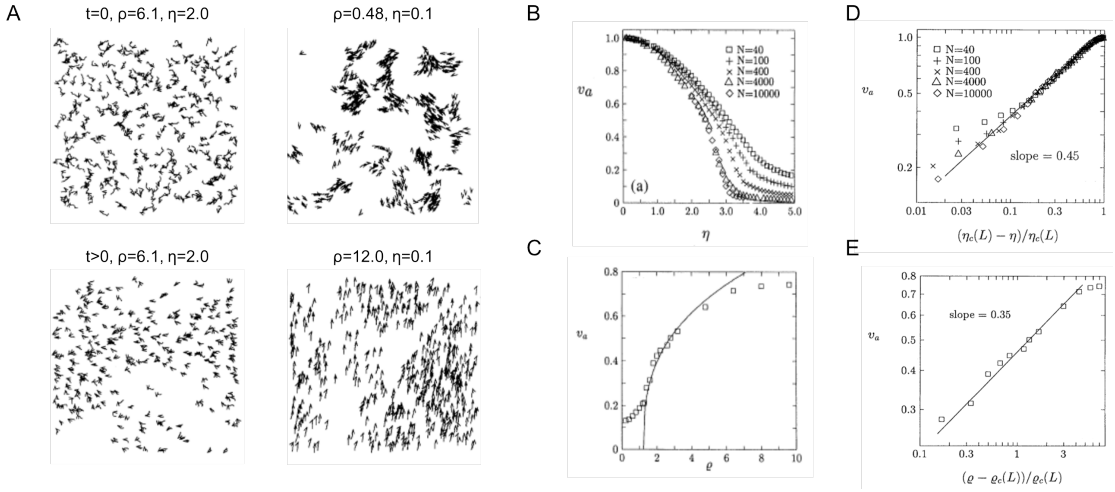

Figure 3.4: The Vicsek model of collective motion. A) Vicsek model simulations for different settings of particle density ρ and noise amplitude η. B) The absolute value of the average normalized velocity, v_a vs the noise η in boxes of various sizes for a fixed density ρ. C) (for $L = 20$) How v_a changes if the noise is kept constant and the density is increased. D) Dependence of $\ln v_a$ on $\ln([\eta_c(L) - \eta]/\eta_c(L))$ and E) $\ln([\rho - \rho_c(L)]/\rho_c(L))$. Panel adapted from Vicsek et al. 1995

3.3.1 Theory of Collective Motion

The basic premise of collective motion is that it is a collective behaviour where individual units of the system interact in a simple (attraction/repulsion) or complex manner (through a combination of simple interactions). The key feature of collective motion is the observation that an individual's action is dominated by the influence of other cells such that it behaves very differently from how it would otherwise do so if it was alone. Quantitatively, collective motion has primarily been described and investigated using a modelling approach based on 'agents', virtual units whose movement are dictated by simple local rules. The simplest model leading to collective behaviour is the model of Vicsek et al. 1995 which assumes only self-propulsion and an effective alignment mechanism between individual units.

An individual i is described by its position $\mathbf{r}_i(t)$ and the angle defining the direction of its velocity $\theta_i(t)$ at time t. The discrete time evolution of one particle is set by two equations. At each time step δt each agent aligns with its neighbours

at a distance r due to a noise term $\eta_i(t)$ and moves at a constant speed v_0.

$$\begin{aligned}\theta_i(t+\delta t) &= \langle\theta_j\rangle_{|r_i-r_j|<r} + \eta_i(t)\\ \mathbf{r}_i(t+\delta t) &= \mathbf{r}_i(t) + v_0\delta t\begin{bmatrix}\cos\theta_i(t)\\ \sin\theta_i(t)\end{bmatrix}\end{aligned} \tag{3.34}$$

where $\langle\cdot\rangle$ denotes the average over all individuals j less than a distance r from i. Given v_0 and r, this model is controlled by two parameters, the density of particles, ρ and the amplitude of the noise, η which governs the local alignment. Varying the two parameters it can be shown that this simple model exhibits a discontinuous phase transition from disordered motion to a large-scale ordered motion that epitomizes collective motion, (Fig.3.4A). This order transition can be captured by an order parameter, here the averaged normalized velocity between 0 and 1, v_a.

$$v_a = \frac{1}{Nv_0}\left|\sum_{i=1}^{N}\mathbf{v}_i\right| \tag{3.35}$$

where N is the total number of particles and $\mathbf{v}_i$ is the particle velocities. This velocity is approximately zero if the direction of individual particles is randomly distributed and $\approx$1 for an ordered direction of particle velocities. At large noise or low density, particles are on average not aligned, (Fig.3.4B,C) and can be described as a disordered gas. At low noise and high density, particles are globally aligned and move in the same direction, and can be interpreted as an ordered liquid, (Fig.3.4B,C). This phase transition is indicated by the collapse of v_a for various N as the size of the system L increases, (Fig.3.4D,E) and $v_a \sim [\eta_c(\rho) - \eta]^\beta$, $v_a \sim [\rho - \rho_c(\eta)]^\delta$ where β, δ are critical exponents and $\eta_c(\rho)$, $\rho_c(\eta)$ are the critical noise and density (as $L \to \infty$). This simple model has been extended to incorporate particles with size and adhesive interactions, (Grégoire and Chaté 2004) and epithelial sheet migration, (Szabo et al. 2006b).

3.3.2 Epithelial Monolayers

The discontinuous phase transition from unordered to ordered collective motion is also observed in epithelial monolayer growth. Using keratocytes from goldfish, Szabo et al. 2006b observed a relatively sharp transition in the averaged normalized

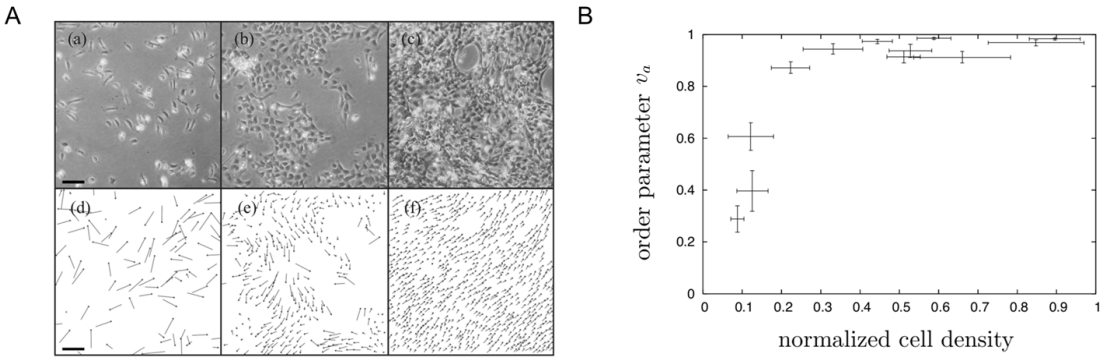

Figure 3.5: Emergence of collective motion in cell monolayers. A) Top: Phase contrast images showing the typical behaviour of cells for three different densities. Bottom: velocity of the cells. Scale bar 50 μm/min. B) Order parameter v_a as a function of normalized cell density. Cell density was normalized with the maximal observed density. Error bars indicate the standard error of the density and order parameter. Panel adapted from Szabo et al. 2006b.

velocity, v_a indicating a transition between random single cell motility to an ordered migration of dense islands of cells, (Fig.3.5A,B). This collective motion behaviour permits the application of PIV to approximate the cellular motion within the confluent sheet. In doing so, Angelini et al. 2011 discovered the presence of dynamic heterogeneities where the fastest cells move in large, multicellular groups where the size of the group grows with increasing cell density. This reported phenomena is reminiscent of particulate systems such as polymers as they approach a glass transition. In highlighting this analogy, Angelini et al. demonstrated techniques typically used to analyse glass-forming systems to extract biological information such as the decomposition of self-diffusion and migration and the cell proliferation rate which would otherwise be difficult to directly measure from the image itself. This work demonstrated that confluent monolayer dynamics under certain conditions might be interrogated using results from condensed matter physics.

Dynamic Structure Factor

In condensed matter physics, the dynamic structure factor is a mathematical function that contains information about inter-particle correlations and their time evolution. It is widely used to characterise material properties. Experimentally the dynamic structure factor is measured directly with inelastic neutron, X-ray, or

light scattering methods, but can be adapted to the analysis of time-lapse images of cell motion to provide dynamical information over a wide range of wavelengths and frequencies, c.f. differential dynamic microscopy (Cerbino and Trappe 2008). Formally, the dynamic structure factor $S(q,\omega)$ is defined as the modulus-squared of the time and space Fourier transform of a dynamic variable such as electron density or neutron density (Sinha 2001). For time-lapse cell images the dynamic variable of interest is the temporal sequence of image pixel intensities, $I(x,y,t)$. Assuming an isotropic sample the dynamic structure factor is

$$S(q,\omega) = \langle |\mathrm{FFT}_{x,y,t}[I(x,y,t)]|^2 \rangle_{\varphi} \tag{3.36}$$

$$S(q) = \langle |\mathrm{FFT}_{x,y,t}[I(x,y,t)]|^2 \rangle_{\varphi,t} \tag{3.37}$$

where FFT denotes the fast fourier transform, $\langle\cdot\rangle_{\varphi}$ indicates an average over the azimuthal angle in 2D q-space and $\langle\cdot\rangle_{\varphi,t}$ indicates an additional average over time. Frequencies and wave vectors are sampled at integer factors of π/L where L is the duration of the set of images or the spatial extent of the image. In particular $S(q,\omega)/S(q)$ can be fitted to the equation of a damped harmonic oscillator (DHO),

$$\frac{S(q,\omega)}{S(q)} = I_0(q)\frac{\frac{1}{2}\Gamma_0(q)}{\omega^2 + (\frac{1}{2}\Gamma_0(q))^2} + I(q)\frac{\Omega(q)\Gamma^2(q)}{[\omega^2 - \Omega^2(q)]^2 + \omega^2\Gamma^2(q)} \tag{3.38}$$

where the first term is a Rayleigh peak, with amplitude $I_0(q)$, width $\Gamma_0(q)$ that quantifies self-diffusivity, the motion due to stochasticity. The second term is a Brillouin peak, with amplitude $I(q)$, width $\Gamma(q)$, peak position $\Omega(q)$ and is essentially a cellular interaction term. Formally it reflects an elastic response to density fluctuations, (Ruocco and Sette 1999). Both these peaks were empirically shown to be essential to describe confluent cell layers, (Angelini et al. 2011).

Transition Cell Density in Confluent Monolayers

A key utility in the decomposition of the dynamic structure factor is the ability to deconvolute and independently assess the stochastic motion due to diffusion (Rayleigh peak) and active motion contributions (Brillouin peak) from cellular processes such as cell division and cell shape changes. The width of the central

peak in the DHO probes the short-time random fluctuations of the cell motion and is shown to be given by $\Gamma_0(q) \sim q^2$, (Angelini et al. 2011) where q is the spatial wave vector. A well defined self-diffusion coefficient for all cell densities is then extracted by averaging over the high q-range

$$D_0 = \langle q^{-2}\Gamma_0(q)\rangle_q \tag{3.39}$$

From this an average distance that relates to diffusion only motion can be computed as a function of a short time interval τ using $\sqrt{D_0/\tau}$ and compared to the average distance from PIV analysis of the phase contrast images which pertains to migration ($100 \leq \tau \leq 200$ min). Angelini et al. 2011 found that both diffusion and migration distances decreased with increasing cell density. At sufficiently large cell densities it is suggested that the diffusion distance becomes longer than the migration density. This convergence defines a transition cell density, σ_g. Below this density cells move greater distances by directed migration than by stochastic motions. Above this density, individual cell motion is predominantly diffusive-like, and cell layer motion is similar to concentrated deformable particles, the fluid-like cell motion becomes increasingly glass-like.

Glassy Behaviour of Confluent Monolayers

Angelini et al. 2011 empirically demonstrate the glass-transition by fitting the diffusivity to the Avramov-Milchev (AM) (Avramov 1998) and the Vogel-Fulcher-Tamman (VFT) equations, two equations commonly used to describe the glass transition in glass-forming colloidal fluids where now the cell density plays the role of the volume fraction.

Avramov-Milchev (AM) equation

$$D_0 = D_{\max} \exp\left[-\varepsilon\left(\frac{\sigma}{\sigma_g}\right)^{\alpha}\right] \tag{3.40}$$

where σ is the cell density, $D_{\max}$ is the diffusivity at zero density, ε is the dimensionless activation energy at the glass-transition density, σ_g is the glass-transition density and α is a fragility parameter.

Vogel-Fulcher-Tamman (VFT) equation

$$D_0 = D_{\max} \exp\left[-K_1\left(\frac{\sigma}{\sigma_g} - K_2\right)\right] \tag{3.41}$$

where σ is the cell density, $D_{\max}$ is the diffusivity at zero density, K_1 and K_2 parameters to be fitted, σ_g is the glass-transition density.

Angelini et al. 2011 show both models fitted equally well to their data revealing that the cell monolayer exhibits the non-Arrhenius relaxation behaviour expected of a moderately fragile glass-forming fluid. Further they extracted $D_{\max} = 30 \pm 4\mu$ m^2 h^{-1} and showed that diffusive-like motions of cells within a confluent layer is more than an order of magnitude slower than the anomalous migration of equivalent isolated MDCK cells, (Dieterich et al. 2008).

Density of States Analysis An additional means of analyzing the information provided by the dynamic structural factor, $S(q,\omega)$ is through the density of states (DOS) which has the potential to better characterise the molecular interactions. In statistical physics the density of states describe the number of states per interval of energy at each energy level available to be occupied. Regarding the monolayer as a 2D system the DOS is given by $n(\omega) = \frac{2q}{q^2_{\max}}\left(\frac{dq}{d\omega}\right)$ where $dq/d\omega$ is the reciprocal of the group velocity, $c = d\Omega(q)/dq$. Because there is no well defined cutoff in the monolayer, such as an intermolecular spacing in crystals, Angelini et al. 2011 analyzed fluctuations in the cell layer down to subcellular length scales where the pixel size was arbitrarily chosen to define the maximum wave vector $q_{\max} = 2\pi/L_{\text{pixel}}$. Interestingly through this analysis they discovered a low frequency cellular process (low ω) and a high frequency cellular process (high ω) corresponding to emergent peaks in the DOS whose scales of motion were on the order of the cell length $l_c \sim \sigma^{-1/2}$, with different wavelengths and temporal scales $\lambda^*_{\text{low}} = 1.19 \pm 0.14 l_c$, $\tau^*_{\text{low}} = \sim$1h to 1.6h, and $\lambda^*_{\text{high}} = 0.39 \pm 0.05 l_c$, $\tau^*_{\text{high}} = \sim$0.4h to 1h. The low frequency mode was attributed to cell body shape fluctuations and the high frequency mode to the characteristic length scale and time scale of motion due to cell division. As cell density increases, the high frequency mode dominated.

Estimating Cell Proliferation from the Glass Transition Perhaps most impressively is that Angelini et al. 2011 was able to recover the cell proliferation rate as a consequence of the glass transition. At the transition cell density, the only motion in the cell layer will be associated with that of cell divisions. Within one division time, all material within each cell is partitioned into two halves of each cell body. By dimensional analysis the length scale of this single-cell motion is $\sim (\frac{1}{2}\sigma_g^{-1})^{1/2}$. The associated time scale using the characteristic diffusion equation is then $\tau_g \sim \frac{1}{2}\sigma_g^{-1}D_g^{-1} = 42.8\text{h}$ where $\sigma_g = 2,800\text{mm}^{-2}$ and $D_g = 4.17 \pm 0.15\mu m^2\text{hr}^{-1}$ is from the fitted VFT equation. For comparison the average cell division time was empirically measured to be 44.1h.

Angelini et al. 2011 conclude that as cell density increases and the transition density σ_g is approached, migration, diffusion, and cell-body deformations all dramatically slow within the cell layer, while the spatial density of cell divisions rises and continues to persist as an increasingly dominant source of high-frequency motion. Whilst this work of Angelini et al. 2011 is elegant and demonstrates that correct interpretation of physical models can be used to quantitatively recover important biological parameters there are several drawbacks that limit this approach for general use. First the use of frequency-based methods places a limit on the imaging frequency due to the Nyquist sampling theorem. Given a finite limit on electronic storage media this effectively prevents the analysis of long-time imaging (days). Second the application of the dynamic structural factor assumes isotropic samples and single particle species which may be approximately satisfied for monolayers but it is unclear how to interpret the results for multiple cell types and the case of epithelial sheet migration where there is an initial gap. Third the derived results are ensemble averages that describe the monolayer and do not retain spatial information. We cannot interrogate areas of abnormal cell division for example or find where in the image a process is happening. Lastly and most important practically, it is difficult for non-specialists to interpret and compute the statistics.

3.3.3 Epithelial Sheet Migration

Compared to the motions of cells within monolayers, the migration of entire sheets has been studied less theoretically and experimentally with less sophisticated computational tools. Two classes of models are prevalent in the literature, partial differential equations based on the Fisher-Kolmogorov equation that yield travelling wave solutions and cellular automata models which describe the local stochastic movement and division of cells. Whilst partial differential equations yield analytically nice models whose behaviour can be systematically interrogated its ability to represent a diverse range of phenomena is more limited as it is not trivial to incorporate local interactions into such an equation. Conversely cellular automata models are extremely flexible and easy to define a complex set of interactions but it is non-trivial to predict and describe emergent global dynamics.

Partial Differential Equations

Epithelial sheet migration has been extensively studied with partial differential equations (PDEs) in the context of a scratch wound assay. Given a confluent layer of cells, a vertical wound is created either chemically or by scratching and the closing of the wound is observed over time. A 'wound' or gap can also be introduced via patterning techniques. It is common to describe this process mathematically in a reaction-diffusion equation called the Fisher-Kolmogorov equation.

$$\frac{\partial c}{\partial t} = D_c \nabla^2 c + \lambda c \left(1 - \frac{c}{k}\right) \tag{3.42}$$

This is Fick's 2nd law of diffusion, $\frac{\partial c}{\partial t} = D_c \nabla^2 c$ where c is the concentration or cell density, t is time with a collective diffusion coefficient D_c that has been augmented with a logistic growth term defined by a growth rate λ and a maximum carrying capacity, k (the maximum cell density possible in the layer). This equation can be shown to exhibit travelling wave solutions corresponding to a stationary density profile propagating with constant speed, $v = \sqrt{4\lambda D_c}$ assuming no-flux boundary conditions. Whilst this model captures to a certain extent the large-scale motion of the propagating wavefront, it cannot readily capture phenomena within the bulk

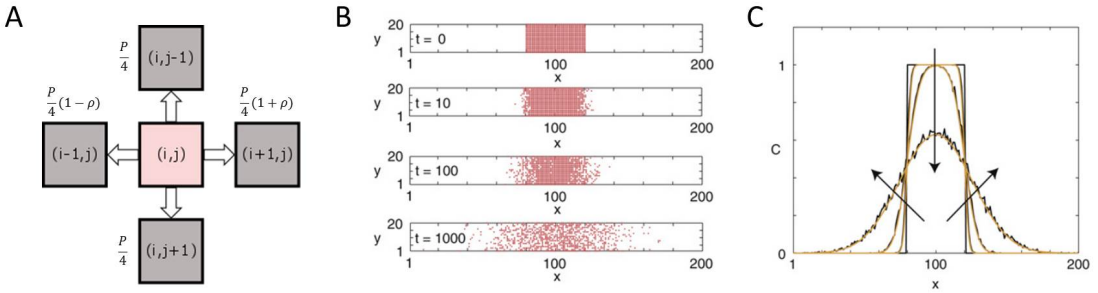

Figure 3.6: Cellular automata model of epithelial sheet migration. A) Local motility rules, P is the probability of moving, ρ is a motility bias where $|\rho| < 1$. B) Simulation with $P = 1$, $\rho = 0$. C) Measured cell density profiles (black) compared to the solution of the continuum solution of the cellular automata (yellow). Adapted from Simpson et al. 2009

of the sheet. For example, the analysis of flow patterns within resting cell sheets reveal the formation of transient swirls of collective rotation reminiscient to that of turbulent fluids. These vortices commonly occur in migration experiments (Angelini et al. 2010). PIV analysis has been used to find that these vortices are on the order of 10 cells large (Angelini et al. 2010; Vedula et al. 2012), with lifetimes of approximately 15 min (Marel et al. 2014b). Increasing cell density increases the swirl correlation length (Angelini et al. 2010) and reduces their overall frequency (Marel et al. 2014b). Another source of local anisotropy is cell division where it has been found that when embedded in connected tissue the cell division axis aligns with the surrounding cellular flow (Marel et al. 2014a). The emergence of tissue scale properties from the properties of individual cells motivates the study of cellular automata models.

Cellular Automata

Cellular automata is an extremely flexible class of modelling which can be used to efficiently investigate the emergent dynamics of a system composed of numerous individual units called 'agents' whose action and interactions with other agents are determined by a specific set of predefined local rules. Perhaps the most famous example of a cellular automata is Conway's Game of Life which gives rise to a shockingly diverse set of behaviour based on only 4 very simple rules. For epithelial sheet migration one of the simplest cellular automata is the exclusion process,

(Schütz and Domany 1993; Evans et al. 1995; Schütz 1997; Simpson et al. 2009) (Fig.3.6). In this model the motility mechanism is implemented on a rectilinear lattice with dimensions $L_x \times L_y$. The lattice discretisation is assumed to be square with spacing Δ. Each site is either empty or occupied by a single agent exclusively. During each time step of length Δt all agents move with probability P. A motile agent spatially indexed by (i, j) steps vertically to $(i, j \pm 1)$ or horizontally to $(i \pm 1, j)$ with a probability $(1 \pm \rho)/4$ where $|\rho| \leq 1$ is a constant which controls the motility bias in the horizontal direction, (Fig.3.6A). If for any particular attempted move, the target site is already occupied, then that move is aborted due to spatial 'exclusion'. It can be shown that the continuum limit of this model yields the following PDE,

$$\frac{\partial c}{\partial t} = D\nabla^2 c - v\frac{\partial}{\partial x}[c(1-c)] \tag{3.43}$$

where the second term is a nonlinear convection term. To solve this equation no flux boundary conditions are usually applied in the horizontal direction with periodic boundary conditions in the vertical direction. Unlike the Fisher-Kolmogorov travelling wave solution, the solution here yields a dispersive solution where the maximum value decreases over time, (Fig.3.6B,C). Naturally such models are overly simplified. In particular it does not model intercellular forces or local alignment and all cells are essentially particles with fixed shapes. As a result modelled 'cells' move individually rather than in cohesive groups as would be expected in a confluent epithelial layer, (Fig.3.6B). Vertex models extends this simple cellular automata to incorporate forces either explicitly or via energy based modelling, (Fig.3.7A) and uses agents operating on non-lattice grids with finite area and volume, (Fletcher et al. 2014; Alt et al. 2017b), Fig.3.7. Non-lattice dynamics permits a more diverse movement set which leads to more realistic cell behaviour for packed epithelial cells, (Fig.3.7B). In particular the explicit force modelling allows for inference of forces which would otherwise be difficult to obtain through empirical measurements, (Alt et al. 2017b). A good example is the modelling of organoid growth where it is challenging to segment the organoid from images or to directly measure the mechanical stresses without perturbing the complex

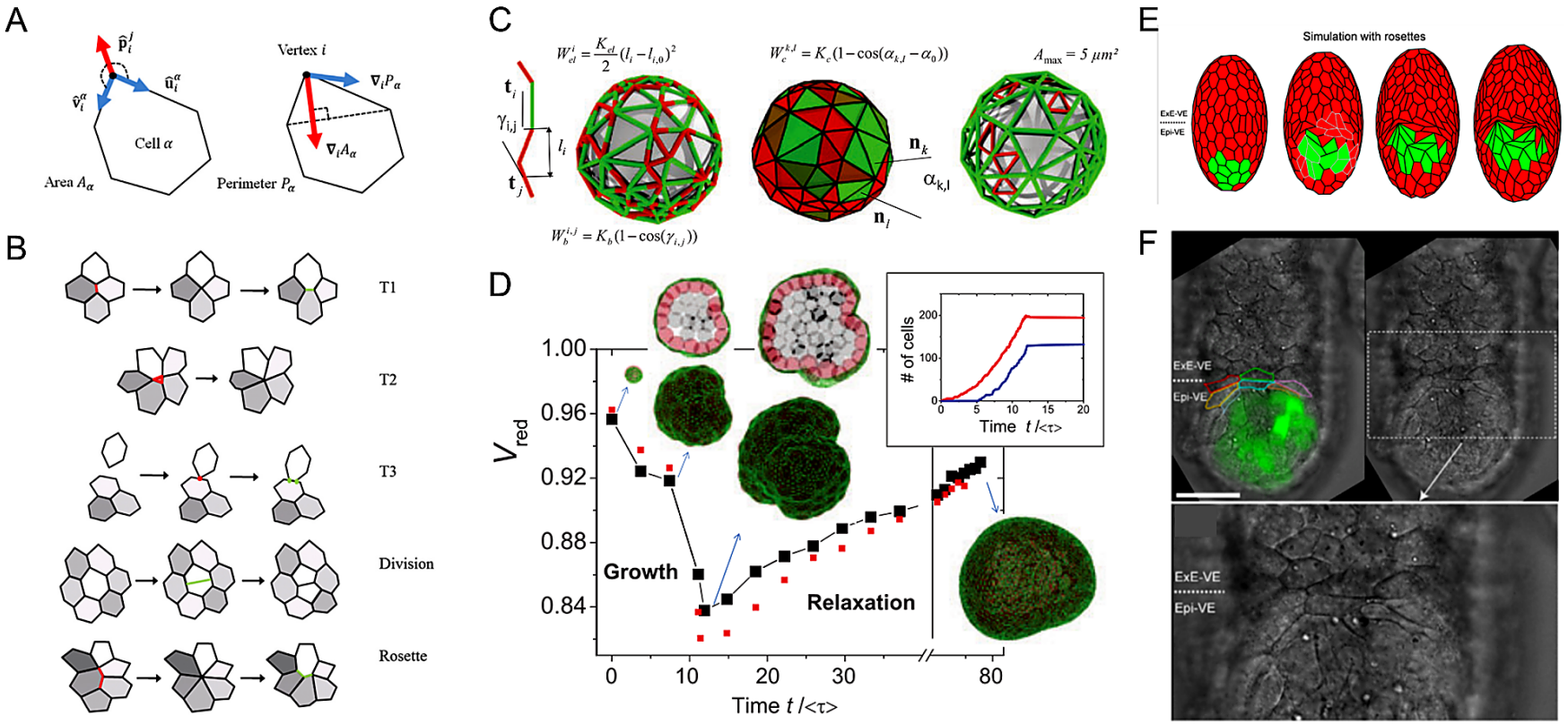

Figure 3.7: Vertex models and its biological application. A) Illustration of key parameters in the explicit force- (left) and energy-based (right) models of vertex mechanics, from Fletcher et al. 2014. B) Schematic diagram of junctional rearrangements, cell division, and cell removal in vertex models, from Fletcher et al. 2014. C) The basal membrane (BM) network, a vertex model incorporating stretching, compression, bending and remodelling for modelling organoids from Buske et al. 2012. D) Simulated dynamics of organoid growth from Buske et al. 2012. E) Cells migrate in a single group when rosettes are allowed to form. Three rosettes are highlighted in grey in the second image from left. At the barrier between the epiblast and extra-embryonic ectoderm, cells form a crescent-shaped group similar to that observed in F) experiment snapshots. Images from Trichas et al. 2012.

morphology of the matrix-embedded organoid, (Fig.3.7C,D). Another example is the modelling of epithelial rearrangement on 3D surfaces. Biological imaging typically does not exhibit good edge resolution and contrast compared to natural images. With 2D image capture of dynamics of a 3D surface for example with embryo development there is also a problem of perspective. Direct 3D modelling incorporating qualitative hypotheses with vertex models can be a good solution that avoids the difficulties of direct quantification, (Fig.3.7E,F).

PIV Analysis and Segmentation Approaches

A popular approach for analysing the motion field of migrating epithelial sheets is the application of PIV, particularly for scratch wound assays, (Milde et al. 2012; Deforet et al. 2012). Most analysis focus on the motion prior to wound closure and quantifies the behaviour using average orientation distributions and speed. The motion field is computed from PIV and temporally averaged over a few frames to obtain the average cellular motion field from which orientation can be analysed, (Tambe et al. 2011). Trajectories are derived to visualize the movement of the wound edge, (Zaritsky et al. 2012a). It is also common to compute the speed of closure of the wound. This is derived by segmenting the epithelial sheets and monitoring the decrease in wound area. Masuzzo et al. 2016 summarise the present state of analysis of wound healing. A notable issue with PIV is the inherent low resolution due to the limit set by the size of the image window. A notable issue with segmentation methods for example to identify collective clusters or to segment foreground and background pixels is that results can be highly variable across time due to noise from background intensity fluctuations (e.g. Ch.6.5.4). The individual disadvantages of pure modelling and imaging based approaches motivates the investigation of hybrid methods.

Hybrid Image Analysis and Modelling Approaches

Hybrid methods that combine pure computer vision and mathematical modelling approaches such as partial differential equations have recently emerged as a potential way to marry ‘empirical’ and ‘theoretical’ approaches to yield more accurate and robust quantification of biophysical parameters that are physically plausible and

is fitted from experimental data. A particular interesting example is the work of Boquet-Pujadas et al. 2017, BioFlow which elegantly demonstrates how to incorporate a theoretical model of the imaged motion here the Navier-Stokes equation into the optical flow equation as a regularization term instead of the usual smoothness constraint (c.f. Eqn.(2.9)) to simultaneously recover pressure, forces and velocities within a moving amoeboid in 2D and 3D.

3.3.4 Motion Interaction Analysis

Having characterised the biological motion parameters of interest, the ultimate aim of automated video analysis is to interpret the motion interaction and understand the motion dynamics within the video. This is also the most important and most difficult task to capture and quantify. It is a higher-level reasoning task that involves analysing the relationship between groups of individual motion trajectories. There are numerous applications that depend on the ability to identify different types of motion from videos for example the detection of anomalous motion in crowds for security surveillance or the identification of novel cellular interactions as a result of an external perturbation such as a genetic mutation or drug treatment. We briefly review some of the existing approaches starting from the simple notion of correlations between trajectories as a means to characterize the level of order in a system. We end by describing some works in computer vision which attempt to identify and automatically group distinct motion patterns in a video to recognise the underlying complex social dynamics.

Cross-Correlation

For two discrete signals, the cross-correlation is defined as

$$(f * g)(t) := \sum_{\tau=-\infty}^{\infty} f^{\dagger}(t) g(t+\tau) \tag{3.44}$$

where $f^{\dagger}$ denotes the complex conjugate of f and τ is the lag time. If $\tau > 0$, $g(t+\tau)$ *leads* $f(t)$. If $\tau < 0$, $g(t+\tau)$ *lags* $f(t)$. This yields a function whose values are unbounded, $[-\infty, \infty]$ quantifying the correlation between two signals of potentially

different lengths. Two signals are positively correlated at time t if $(f * g)(t) > 0$ and negatively correlated otherwise. Numerous applications utilise the cross-correlation for example matching applications that search for a similar looking subimage in a second image or for estimating the time delay between two acoustic signals for speech synchronisation. For comparison of signals which may differ in the absolute values it is useful to define the zero-normalised cross-correlation (ZNCC).

$$\mathrm{ZNCC}(t) := \frac{1}{n} \sum_{\tau=-\Delta T}^{\Delta T} \left(\frac{f^{\dagger}(t) - \bar{f^{\dagger}}}{\sigma_f^{\dagger}} \right) \left(\frac{g(t+\tau) - \bar{g}}{\sigma_g} \right) \tag{3.45}$$

where ΔT is the maximum lag. For motion, the most common cross-correlation measure is that of the velocity cross-correlation where $f = V_i$ and $g = V_j$ are the velocity trajectories of two different motion trajectories i and j.

$$V_i = \{v_i^1, v_i^2, ..., v_i^t | t \in [1, T]\} \tag{3.46}$$

Computing the pairwise velocity cross-correlation between individual trajectories belonging to separate entities, interaction graphs can be deduced by appropriate thresholding. Shishika et al. 2014 analysed the swarm behaviour of mosquitos in this manner and showed that males form synchronized subgroups whose size and membership changes rapidly. This quantity is also referred to as the *directional correlation function*. Another major application is in studies of flocking of birds and fish where the time delay for which the cross-correlation attains the maximum is used to determine the leader-follower relationship (e.g Nagy et al. 2010). A negative value indicates that individual i lags j and vice versa. Another cross-correlation we use in this thesis is when the signals are the positional trajectories of an individual i and j, $f = \mathcal{T}_i$ and $g = \mathcal{T}_j$. We term this the track similarity and it measures the similarity in the shape of two trajectories whilst also factoring in the spatial distance between them.

Spatial-Correlation

Whereas the velocity cross-correlation aims to measure temporal correlations in time, the spatial correlation aims to assess the spatial length-scale over which there

is correlation between individuals i and j.

$$C(r) = \frac{1}{c_0} \frac{\sum_{ij} v_i \cdot v_j \delta(r - r_{ij})}{\sum_{ij} \delta(r - r_{ij})} \tag{3.47}$$

where the velocity vector v_i is obtained after subtracting the global mean velocity from each vector such that the mean $\mathbb{E}[v_i] = 0$, (c.f. Cavagna et al. 2010). c_0 is a normalization factor such that $C(r = 0) = 1$. Generally the correlation decreases with increasing distance r. Beyond a certain distance the correlation tends to 0, indicating the absence of correlation. It is usual to summarise the decreasing trend of this curve using the concept of a correlation length, ζ which for exponentially decaying curves is taken to be the decay constant, or in Cavagna et al. 2010 the zero crossing point of the correlation function for a flock. This measure can be a powerful way to describe the approximate size over which individuals are correlated in motion. Using this measure, Cavagna et al. 2010 surprisingly found that the correlation length ζ for starling flocks does not exhibit a constant value either in units of number of birds or in units of metres but grows linearly with the size of the flock.

Measuring Collectiveness

The order parameter given by the average normalised velocity, v_a defined by Eqn.(3.35) is the most popular and prevalent method for quantifying the onset of collective motion. However there are several key shortcomings of the approach that limits its applications to a broad range of videos. There is no notion of memory for example thus it is not well defined for particles that move but then stop and it is unstable in the presence of large noise. Compared to the tasks of segmentation few works explicitly address the need for a better collective metric which can be used for soft grouping. One of the more appealing works is that of Zhou et al. 2013 whom propose the notion of collective manifolds, spatiotemporal coherent structures that emerge from the collective movements of constituent individuals in crowds. To this end they appeal to two key properties of collective manifolds:

- *Behaviour consistency in neighbourhoods:* neighbouring constituent individuals exhibit consistent behaviours.

- *Global consistency among non-neighbours:* although constituent individuals at a distance may have different behaviours, they can be correlated by behaviour similarity through intermediate individuals in neighbourhoods on the manifold.

Crowd collectiveness is then framed as a measure of the holistic behaviour consistency of the collective manifold. Zhou et al. 2013 proposed to quantify this consistency in a bottom-up manner using the notion of paths on the collective manifold: aggregating behaviour consistency in a local neighbourhood through the spatial proximity of neighbours up to the entire video.

Behaviour Consistency in a Neighbourhood. When an individual j is in the neighbourhood of i, $j \in \mathcal{N}_i$ at time t the similarity is defined as

$$w_t(i,j) = \max(C_t(i,j), 0), \tag{3.48}$$

$$C_t(i,j) = \frac{\mathbf{v}_i \cdot \mathbf{v}_j}{||\,\mathbf{v}_i\,||_2 ||\,\mathbf{v}_j\,||_2} \tag{3.49}$$

where $C_t(i,j)$ is the cosine similarity between two vectors at t between i and j. $\mathcal{N}$ is the K-nearest neighbourhood. $w_t(i,j) \in [0,1]$ measures an individual's behaviour consistency in its neighbourhood. The above is not as robust when two individuals are not in a neighbourhood therefore the *path similarity* is used to construct a more robust estimate of behaviour consistency, the l-path similarity, $\nu_i(i,j)$. Let $\mathbf{W}$ be the weighted adjacency matrix of the graph where the edge weight is $w_t(i,j)$, the similarity between two individuals i and j in its neighbourhood defined by Eqn.(3.49). A path from individual $i \to j$ of length l through individuals $p_0, p_1, ..., p_l$ is denoted $\gamma_l = \{p_0 \to p_1 \to \cdots p_l\}$, $(p_0 = i,\ p_l = j)$. The path similarity on a specific path is denoted γ_l and is defined

$$\nu_{\gamma_l} = \Pi_{k=0}^{l} w_t(p_k, p_{k+1}) \tag{3.50}$$

There can be more than one path of length l between i and j. Let the set $\mathcal{P}_l$ contain all the paths of length l between i and j, then the l-path similarity is defined

$$\nu_l(i,j) = \sum_{\gamma_l \in \mathcal{P}_l} \nu_{\gamma_l}(i,j) \tag{3.51}$$

Collectiveness at l-Path Scale. The individual collectiveness of an individual i at l-path scale is defined as

$$\phi_l(i) = \sum_{j \in \mathcal{C}} \nu_l(i,j) = [\mathbf{W}^l \mathbf{e}]_i \tag{3.52}$$

$\mathcal{C}$ is the set containing all the individuals in the crowd and $\mathbf{e}$ is a vector with all elements as 1, $[\cdot]_i$ is the i-th element of a vector. The crowd collectiveness at l-path scale is defined as the mean of the individual collectiveness

$$\Phi_l = \frac{1}{|\mathcal{C}|} \mathbf{e}^\intercal \mathbf{W}^l \mathbf{e} \tag{3.53}$$

Regularizing Collectiveness over All Scales. Ideally we would like to integrate the individual and crowd collectiveness over all the path scales, $l = \{1, 2, ..., \infty\}$. However the exponential growth of ϕ_l and Φ_l prevents direct summation approaches. Zhou et al. 2013 therefore propose the use of generating function regularization to assign a meaningful value for the sum of a possibly divergent series, (Knuth 1997). For the l-path similarities they use

$$\tau_{i,j} = \sum_{l=1}^{\infty} z^l \nu_l(i,j) \tag{3.54}$$

where z is a real-valued regularization factor that cancels the exponential growth of ϕ_i, Φ_l with l. z^l acts like a 'weight' for the l-path similarity at different l. It can be shown that the individual collectiveness from the generating function regularization on all the path similarities can then be written as

$$\phi(i) = \sum_{l=1}^{\infty} z^l \phi_l(i) = [\mathbf{Z}\mathbf{e}]_i \tag{3.55}$$

where $\mathbf{Z} = (\mathbf{I} - z\mathbf{W})^{-1} - \mathbf{I}$. The crowd collectivess of a crowd system $\mathcal{C}$, defined as the mean of all the individual collectiveness is then in closed form,

$$\begin{aligned} \mathbf{\Phi} &= \frac{1}{|\mathcal{C}|} \sum_{i=1}^{|\mathcal{C}|} \phi(i) \\ &= \frac{1}{|\mathcal{C}|} \mathbf{e}^\intercal \mathbf{Z} \mathbf{e} \end{aligned} \tag{3.56}$$

Convergence Condition. For the proposed measure there are two parameters z and K for computing the collectiveness. K is the topological range of the neighbourhood and z is the regularisation constant. Tuning z and K controls the sensitivity of collectiveness. The authors recommend $\frac{0.4}{K} < z < \frac{0.8}{K}$ for stability.

To apply this method, one would usually compute the dense optical flow and remove all points that do not move sufficiently. One of the drawbacks of this is the presence of oversampling of large objects in the field of view: with uniform sampling large objects naturally possess more feature points than small objects. The number of feature points therefore does not reflect the real number of individuals in the scene. Consequently collectiveness measured using the topological relations between points can be skewed. Li et al. 2016a proposes an initial point selection procedure using a superpixel approach. Additionally they propose spatiotemporal neighbourhoods. Manifold ranking, (Zhou et al. 2004) is used to define a similar crowd collectiveness quantity

$$\mathbf{\Phi} = \frac{1}{|\mathcal{C}|}\mathbf{e}^{\intercal}\mathbf{Z}_t\mathbf{e} \tag{3.57}$$

$$\mathbf{Z}_t = (1-\alpha)(\mathbf{D} - \alpha\mathbf{W})^{-1} \tag{3.58}$$

where $\mathbf{D}$ is a matrix that encodes additionally the physical euclidean distance between neighbours in the neighbourhood and $\mathbf{W}$ is an affinity matrix that includes the temporal connection. An optimal number of superpixels was noted for best performance. Too many superpixels allow redundant points to be kept whilst too little leads to the discarding of potentially informative points.

Social Force Models

The above measures characterise the order of the individuals in a system and computer vision models may be able to automatically identify distinct motion groups based on these measures but they do not explicitly capture the inter-relationships between different motion patterns or describe how they might change. For example, is an individual following an individual? Is a group forming? Is

a group dispersing? Is a group meeting another group? Having recognised such primitives how do we understand the entire scene? What is the motivation behind an action? One of the earliest attempts to resolve such questions in a non-scene specific manner is through the social force model proposed by Helbing and Molnar 1995. In this model each of N pedestrians i with mass of m_i changes his/her velocity v_i as a result of an actual force F_a which consists of (1) a personal desire force F_p and (2) an interaction force F_{int}.

$$m_i \frac{dv_i}{dt} = F_a = F_p + F_{int} \tag{3.59}$$

People in crowds generally seek certain goals and destinations in the environment. Each pedestrian therefore have a desired direction and velocity v_i^p. The crowd meanwhile limits individual movement and the actual motion of a pedestrian is v_i which differs from their desired velocity. Over time, individuals tend to approach their desired velocity v_i^p based on the personal desire force

$$F_p = \frac{1}{\tau}(v_i^p - v_i) \tag{3.60}$$

where τ is the temporal relaxation parameter. The interaction force F_{int} consists of the repulsive and attractive force F_{ped} based on the psychological tendency to keep a social distance between pedestrians and an environment force F_w to avoid hitting walls, buildings and other obstacles.

$$F_{int} = F_{ped} + F_w \tag{3.61}$$

Naturally we would also expect pedestrians to keep small distances with people they are related to or attracted to whilst keeping far away from discomforting individuals or environments. Such observations can be explicitly modelled with potential fields, (c.f. Helbing and Molnar 1995) and included accordingly as additional terms in Eqn.(3.59). Individuals with high interaction force have a higher probability of belonging to the same social group with the same collective dynamics. Changes in the interaction forces in time determines the ongoing evolution of the crowd behaviour.

Mehran et al. 2009 showed how to simply construct a force flow for a video based on particle advection and optical flow for the generalised social force model that considers the effect of *panic*, where herding (crowd grouping) behaviours appear in events such as escaping from a hazardous incident. For this model, the personal desire velocity v_i^p is replaced with

$$v_i^q = (1 - p_i)v_i^p + p_i\langle v_i^c \rangle \tag{3.62}$$

where p_i is the panic weight parameter and $\langle v_c \rangle$ is the average velocity of the neighbouring pedestrians. As $p_i \to 0$ pedestrian i exhibits individualistic behaviours. If $p_i \to 1$ pedestrian i exhibits herding behaviours. The generalized social force model is then

$$m_i \frac{dv_i}{dt} = F_a = \frac{1}{\tau}(v_i^q - v_i) + F_{int} \tag{3.63}$$

To avoid image segmentation, particles are densely seeded and advected with optical flow, O. The actual velocity of the particle is denoted $v_i = O_{ave}(x_i, y_i)$. $O_{ave}(x_i, y_i)$ is the averaged optical flow averaged over a small spatial window around (x_i, y_i) and across a few frames. Substituting the desired velocity of the particle in the generalised social force model, Eqn.(3.63) is then

$$v_i^q = (1 - p_i)O(x_i, y_i) + p_i O_{ave}(x_i, y_i) \tag{3.64}$$

where $O(x_i, y_i)$ is the raw computed dense optical flow at (x_i, y_i). The interaction force F_{int} is estimated from Eqn.(3.63) assuming that each object in the scene is of similar size such that $m_i = 1$.

$$F_{int} = -\frac{1}{\tau}(v_i^q - v_i) + \frac{dv_i}{dt} \tag{3.65}$$

To detect panic, compute F_{int} with $p_i = 0$. Any high magnitude interaction force consequently relates to activities that differ from the crowd movement. For each frame this produces a force flow field. Mehran et al. 2009 subsequently trained a latent dirichlet allocation model using only the 'normal' videos to classify normal and abnormal video frames.

One of the limitations of Mehran et al. 2009 is no explicit mechanism to identify more complex behaviours. Scovanner and Tappen 2009 extend the ideas of the original social force model to propose an energy based continuous optimization approach to learn the pedestrian's movement using gradient descent. However this approach is limited to the prediction of a single individual's movement and as input requires the individual's initial position, velocity and the locations of the obstacles in its path. Therefore an accurate object detection algorithm would additionally be required for automated analysis in practice.

Group-Level Scenario Recognition

By explicitly modelling the motivations of individuals, social force models can provide interpretable solutions to the movement of individuals in a scene. As shown by Scovanner and Tappen 2009 it can help to enforce realistic continuous solutions to the tracking problem or to locate regions of anomalies that deviate from the model expectations. Nevertheless they do not address the issue of learning and recognizing higher level dynamic interactions. This is a very challenging task and few papers explicit address it. Chang et al. 2011 present a neat systematic solution which effectively combine the advantages of soft grouping of motion in a weighted connection graph and probabilistic models for motion interpretation both at the individual and group level. We defer the detailed mathematical specifics to the paper and present the key ideas. Given a person detection algorithm and the use of Kalman filters for individual person tracking the proposed framework can be divided into two components.

Soft probabilistic grouping. Given a set of individuals and their respective movement tracks in time, a path-based group connectivity matrix, $\mathbf{W}$ is constructed where an edge, w_{ij} connecting two individuals i and j can be interpreted as a

Group scenario	Probabilities for track i, or between tracks (i,j)
Group formation	$p_g^f(i) = \text{sigmoid}(y_g^f, 1, 0.2), y_g^f = \sum_{\forall j \neq i} p_c^\pi(i,j;t) \cdot [1 - p_c^\pi(i,j;t_p)] \cdot \max(h(i), h(j))$
Group dispersion	$p_g^d(i) = \text{sigmoid}(y_g^d, 1, 0.2), y_g^d = \sum_{\forall j \neq i} p_c^\pi(i,j;t_p) \cdot [1 - p_c^\pi(i,j;t)] \cdot \max(h(i), h(j))$
Stable group	$p_g^s(i) = 1 - p_g^f(i) - p_g^d(i)$
Loitering group	$p_g^l(i) = 1 - \prod_{\forall j} \{1 - p_c^\pi(i,j) p^l(j)\}$
Stable loitering group	$p_g^{sl}(i) = p_g^s(i) p_g^l(i)$
Distinct groups	$p_g^\delta(i,j) = \prod_{\forall k} \{1 - \max(p_c^\pi(i,k)p_c^\pi(k,j), p_c^\pi(j,k)p_c^\pi(k,i))\}$
Close-by groups	$p_g^c(i,j;t) = 1 - \sum_{k \neq i,j} [1 - p^c(i,k;t)] \cdot [1 - p^c(k,j;t)]$
Group meeting	$p_g^{meet}(i,j) = 1 - \prod_{t=t_0 \text{ to } t_f} \{1 - p_g^c(i,j;t)\}$
Group following	$p_g^{flw}(i,j) = p_g^\delta(i,j) \cdot [1 - \prod_k \{p_g^\delta(i,k) + [1 - p_g^\delta(i,k)] \cdot [1 - p^{flw}(k,j)]\}]$
Group chasing	$p_g^{chs}(i,j) = p_g^\delta(i,j) \cdot [1 - \prod_k \{p_g^\delta(i,k) + [1 - p_g^\delta(i,k)] \cdot [1 - p^{chs}(k,j)]\}]$

Figure 3.8: Probabilistic rules for group-level scenario recognition, from Chang et al. 2011.

probability, p_c^π with value between 0 and 1.

$$w_{ij} = p_c^\pi$$
$$p_c^\pi(\{i \text{ and } j \text{ are connected via } i_0,...,i_N\}) = p_c^p(i,i_0)\left[\Pi_{k=0}^{N-1} p_c^p(i_k, i_{k+1})\right] p_c^p(i_N, j) \tag{3.66}$$

$p_c^p(i,j,t)$ is a pairwise affinity that incorporates the track history.

$$p_c^p(i,j,t) = \omega_1 p_c^{inst}(t) + \omega_2 \frac{\sum_{t_i \in T} p_c^{inst}(t_i)}{|t_i \in T|}, \quad \omega_1 + \omega_2 = 1$$
$$p_c^{inst}(i,j) = f(d_{ij}, \phi_{ij}, ||v_i||, ||v_j||) \tag{3.67}$$

where T is a time window, ω_1, ω_2 adjusts the weights between the current status and the entire window history. $p_c^{inst}(i,j)$ is an instantaneous affinity measure for time t and is a function of d_{ij} the euclidean distance between two individuals located at position $\mathbf{r}_i$ and $\mathbf{r}_j$ respectively, ϕ_{ij} the angle between i's velocity vector and the relative position vector between i and j with respect to person j and the respective speeds of i and j.

Probabilistic Motion Analysis A key insight of Chang et al. 2011 is the proposal of prototypical common interaction rules based on the probabilistic

grouping that can be described mathematically and evaluated based on the pairwise path connectivity between individuals, p_c^{π} and $h(i)$, the track healthiness which depends on the Kalman filter covariance and track lifetime. Rules for recognising group scenarios are shown in Fig.3.8. Critically, one can identify a combination of the different prototypes such as *stable loitering distinct groups* by combining probabilities according to the standard rules of probability. Similar prototypes can also be proposed for individual person motion analysis with similar combination rules. In addition the algorithm can be run in real-time. A follow-up work by Zhang et al. 2012 using derived features of the connectivity matrix and machine learning reports improved classification accuracy over the pure rule-based method of Chang et al. 2011.

3.4 Summary and Conclusions

- Cellular motion events can be captured and modelled at three different levels:
  - Collection of individual particles - Microscopic theory of transport
  - Local particle motion divergences - Lagrangian theory of motion
  - The emergent global phenomena - Collective motion which utilises multiple descriptions
- Different specialised methods are currently used to analyse different cellular motion videos with limited holistic characterisation of behaviour across the different levels.
- PIV approach is widely used for biological motion extraction but downstream analysis is limited and do not systematically account for local and global phenomena.
- Graphs have been shown to be a powerful way to capture local relationships between cells and form the basis for state-of-the-art measures of collectiveness and for analysing social relationships between individuals
- Graphs also enable probabilistic grouping of local motion and probabilistic recognition of the relationships between individuals to analyse the motion interaction between distinct entities

4

Motion Sensing Superpixels

Contents

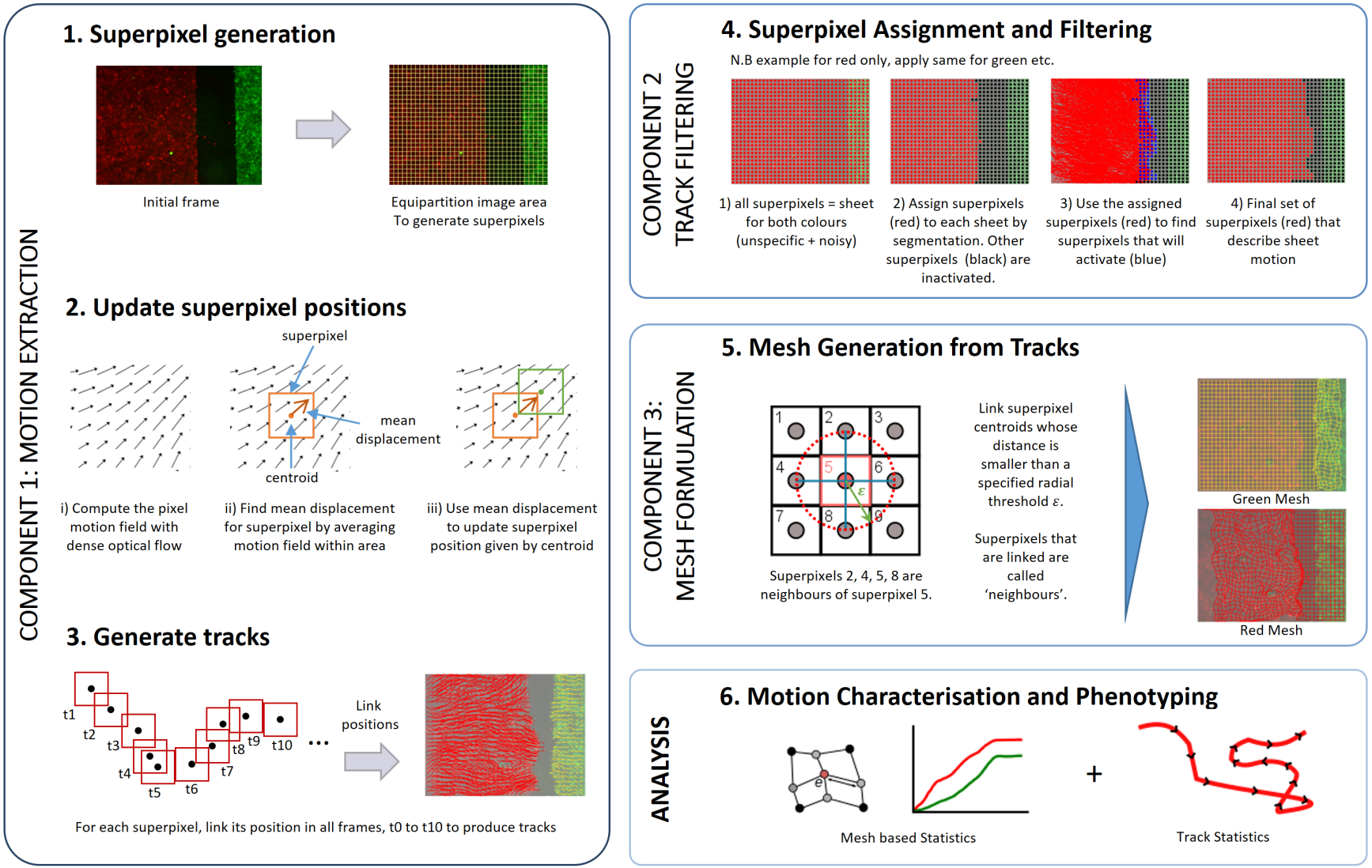

Figure 4.1: Overview schematic of MOtion SEnsing Superpixels (MOSES) framework for motion tracking and analysis. Motion extraction from a video is performed in steps 1-3; filtering of tracks in step 4; motion analysis is based on a mesh formulation (step 5), which can be combined with track statistics to quantitatively analyse motion (step 6).

4.1 Introduction

Motion Sensing Superpixels (MOSES) was developed for the robust quantification and analysis of tissue-level dynamics. The primary aim of MOSES is to provide a systematic method that does not rely on lengthy data specific optimization of multiple parameters, specific image acquisition requirements or complex model fitting. To this end the proposed workflow aims to contain a minimal number of free parameters that the user needs to specify. Wherever possible we avoid the use of image segmentation and machine learning techniques that may overfit to biases present in the data and consequently negatively affect generalisation to other datasets or experimental conditions. The latter is a crucial consideration. Whilst current deep learning approaches have demonstrated state-of-the-art off the shelf performance in classification tasks (Sharif Razavian et al. 2014), reuse of trained networks even for similar objects from other datasets necessitates a degree of fine-tuning to guarantee performance, (Yosinski et al. 2014) and they prone to be 'fooled'. Formally they are susceptible to adversarial attack where the network produces high confidence classification for unrealistic images (Nguyen et al. 2015) and unoticeable modifications of the original image can induce a completely different classification (Goodfellow et al. 2014b). This re-training can be limiting in general for biomedical imaging where the availability of data particularly labelled datasets is a premium and raw data may not readily be crowd-sourced for annotation due to patient ethics and the need for domain-specific expertise. Furthermore deep learning features can be difficult to interpret biologically and there is no guarantee of generalising the learnt features beyond the task for which it was trained on. This phenomenon is sometimes referred to as catastrophic forgetting, (Kirkpatrick et al. 2017; Shin et al. 2017). Particularly illuminating is the findings by Shin et al. 2017. First they trained a neural network to classify the handwriting dataset with no background, MNIST (LeCun et al. 1998) with near 100% accuracy but on a related dataset composed of numbers on a coloured background, SVHN (Netzer et al. 2011), with no re-training the classification accuracy was only 75%. Even if the network was trained on the more difficult SVHN first (90% accuracy), the accuracy was

still poor when applied to MNIST (65%). Thus despite impressive classification accuracy the network does not 'understand' the concept of numbers. At the roots, the primary unsolved problem is that of dataset bias (Torralba and Efros 2011) and the inherent assumption the application dataset is fixed and has the same statistical properties as the training set. In the former due to practical limitations in data collection and diversity, individual datasets exhibit a unique signature that can be distinguished with above chance probability even with shallow classifiers such as support vector machines (SVM) (Hearst et al. 1998). One might expect this to be remedied with a much larger dataset. Google Research recently tested the effect of dataset size. Taking the best classification models from ImageNet, a dataset with >10 million images and applying it to Google's internal image dataset JFT-300M, composed of >375 million images, Sun et al. 2017 found no saturation in performance levels. The more data the better the network performance. 300 million is clearly at present an astronomical figure for labelled image collection from a specialised discipline such as biomedical imaging. In the latter, the assumption of static datasets may not hold in practice for example when the experimenter switches experiental protocol or switches the instrument for data acquisition. Lastly most problems cannot be readily cast as a differentiable problem that is suitable for end-to-end deep learning, see the excellent blog of Francois Chollet (Chollet 2017a; Chollet 2017b).

Despite these drawbacks it is undeniable that deep learning approaches possesses many attractive properties when considering scaling up computation in production and has had huge impact in real-world applications. Taking advantage of modern GPU computing, deep networks can be incredibly fast and by exploiting the ability to learn features directly from data exhibits higher accuracy (increasing with the amount of data) than equivalent hand-crafted approaches. For example, by switching to deep natural lanuage processing networks Google has significantly improved the quality of Google Translate (Research 2016; Wu et al. 2016). To realise these benefits, careful design is required. The most successful and popular network architectures

are all inspired by the design of robust hand-crafted pipelines on which they base the architecture design before making it suitable for end-to-end learning. For example VGG19 (Simonyan and Zisserman 2014b) for image classification makes multi-scale filter banks learnable, Faster R-CNN (Ren et al. 2015) makes region proposals and classification learnable, FlowNet2.0 (Ilg et al. 2016) mimics the optical flow optimization problem with neural networks and Sfm-net (Vijayanarasimhan et al. 2017) casts camera geometry into an end-to-end framework to recover 3D structure from videos.

Consequently the design of MOSES is motivated by techniques from established motion models (Ch.3) whilst at the same time aims to facilitate extraction of features for data-driven machine learning approaches. In this thesis our aim is to first produce a computational framework that delivers sound baseline performance and generalisation even on small datasets ('jack-of-all-trades') before exploring in future work (c.f. Ch.7.4.2) the possibilities of enhancing the existing implementation with end-to-end training capabilities for more efficient computation and the delivery of optimised dataset-specific results ('possibility to train specialised experts').

4.2 General Workflow

4.2.1 Motion Extraction

Due to the inherently large variations in cell appearance and cell density, tracking of individual cells is unreliable. Instead we build MOSES on agnostic global motion extraction using robust dense optical flow similar to the particle image velocimetry (PIV) approach for analysing monolayers and migrating epithelial fronts (e.g. Das et al. 2015) subsampling the video motion without image segmentation, (Fig.4.1). The initial frame is divided into a specified number of regular regions called superpixels; each superpixel is larger than one image pixel (Achanta et al. 2012) (step 1, Fig.4.1). The centroid positions of these superpixels is then updated frame-to-frame, according to the averaged dense optical flow (Farnebäck 2003) displacement inside the area encompassed by the superpixel (step 2, Fig.4.1). A

superpixel track is formed by collecting the centroid positions of a superpixel at all time points (step 3, Fig.4.1). The collection of all superpixel tracks in the video summarises all the motion content present. The number of superpixels used should be chosen to sufficiently subsample the spatial area of the image to capture all spatial variations in motion. For multi-channel videos, we track each channel independently in the described manner. The resultant tracks are not merged by design choice. In biology it is standard practice to visualise different cell types in individual image channels using different fluorescent stains. In this thesis only fixed size square or rectangular superpixels were used for computational efficiency as controlled through the compactness parameter of the SLIC algorithm (Achanta et al. 2012). More generally the superpixels used can be of any shape and can be produced by any other suitable superpixel algorithm e.g. Felzenszwalb and Huttenlocher 2004. Superpixel size may also be allowed to deform in time according to the computed optical flow.

4.2.2 Track Filtering and Post-processing

As the superpixel tracks capture all the motion in the video, some tracks may not be relevant to the phenomenon of interest. Therefore, the second component of MOSES is the post-filtering of tracks (step 4, Fig.4.1) and is currently specific to the experimental setup being analysed. In this step, superpixels are assigned to cover the entire dynamic motion for each 'object' of interest. To assign superpixels to an object, each object is first segmented from the initial video frame to produce a binary image where pixels have value 1 if belonging to the object or 0 otherwise. A superpixel belongs to an object if its centroid lies in a pixel whose value is 1 (red centroids in Fig.4.1 step 4, 2nd image) in the respective binary image, otherwise it does not initially belong to an object (black centroids in Fig.4.1 step 4, 2nd image). To improve motion dynamics estimation for the entire movement of a object based on its assigned initial superpixels, the initial superpixels covering the object is subsequently propagated along its trajectory through time. Any superpixels that did not initially belong to the object (or inactive superpixels) becomes 'activated' and its track considered in addition to the initial superpixels if

they come to lie within the combined spatial area covered by the initial superpixels during the propagation. This is illustrated in Fig.4.1 step 4, 3rd image where the 'activated' superpixels are marked with blue centroids. The final set of superpixel tracks that specifically describe the motion of the object is the union of initial and 'activated' superpixel tracks.

4.2.3 Representing motion as dynamic meshes

The third component of MOSES is a unique approach to motion analysis, in which we generate a 'mesh' from the superpixel tracks (step 5, Fig.4.1). After filtering, each superpixel track is linked to its neighbouring superpixel tracks, according to a euclidean distance cut-off. It is not fitted through a complex optimization routine. This transforms the independent trajectories of each superpixel into a dynamic mesh that naturally assimilates the local collective dynamics. For multichannel videos, a separate mesh is produced for each colour channel. Multiple different meshes can be constructed in such a manner for different purposes, (Ch.4.3.2). This mesh formalism is a key property of MOSES. Because the mesh is a representation of motion that is built from the local spatial relationships of superpixels, robust measures of motion can be derived whilst naturally taking into account the social relationships of cells (c.f. Ch.3.3.4). In the next chapter we will see how the mesh approach also provides a consistent framework to derive video motion signatures to facilitate visual phenotypic comparisons of cellular motion on a 2D plot.

4.2.4 Motion characterisation and phenotyping

MOSES combines mesh-based and familiar single trajectory statistics such as those in Meijering et al. 2012 together to provide a comprehensive, rigorous characterisation of cellular motion from the microscale statistics of cell speed to the macroscale statistics of epithelial sheet interaction and boundary formation by the systematic assimilation of information through the mesh.

4.3 Dynamic Superpixel Meshes

We mathematically describe our mesh formulation of motion based on superpixels. In doing so we introduce notation to precisely define concepts that will be used throughout the thesis.

4.3.1 Definition and Notation

Superpixels and superpixel area. A superpixel i denoted S_i is defined as a geometric region with a centre (x, y) position (centroid), $c_i(t) = [x_i, y_i, t]$ at time t and has an associated spatial area ω_i^t which can be specified using a distance function relative to its centroid e.g. for time $t = 0$, we can define the superpixel area associated with S_i to encompass all pixels with (x, y) closest to the centroid of S_i.

$$\omega_i^0 := \{(x, y) | d_i(x, y) < d_j(x, y) \quad \text{for all } j \neq i\} \ , \tag{4.1}$$

where d_i is a distance function. For square superpixels this is the Manhattan distance where $|\cdot|$ denotes the absolute value.

$$d_i(x, y) := |x - x_i| + |y - y_i| \tag{4.2}$$

For simplicity and computational efficiency to enable high-throughput video analysis, square superpixels are used throughout this thesis and are efficiently implemented in this thesis by setting a high compactness parameter (=10) using SLIC, (Achanta et al. 2008). Naturally other superpixel assignments and areas are possible resulting in different definitions of area ω_i^0, e.g. (Comaniciu and Meer 2002; Felzenszwalb and Huttenlocher 2004) which may suit particular applications. Our mesh formulation is agnostic of the specific superpixel extraction method. We additionally choose to preserve the initial superpixel area for $t > 0$ when computing the average displacement to update superpixel centroid positions for motion extraction. Using square superpixels, we then define the average superpixel width as $w_s = \sqrt{\mathbb{E}[\omega_i^0]}$ where $\mathbb{E}$ denotes the mean.

For analysis, it is easier to consider for $t \geq 0$ the disk area specified by a constant distance cut-off, d_c where d_c can be set using prior assumptions. Assuming for example that each of our square superpixel area can be approximated by a 2D Gaussian, $\mathcal{N}(\mu, \mathbf{\Sigma})$ with a mean position described by the superpixel centroid, $\mu = (x_i, y_i)$ and a covariance matrix, $\Sigma = \begin{bmatrix} w_s & 0 \\ 0 & w_s \end{bmatrix}$. We can then set d_c to be multiples of the effective standard deviation w_s. Usually we set the multiple to be 2 (to include 95% of the distribution). In practice we have used multiples from 1-5 and found the result to not be sensitive to this setting.

$$\omega_i^0 := \{(x, y) | d_i(x, y) < d_c\} \tag{4.3}$$

We denote the set of all superpixels for a video with $\mathcal{S}$.

Temporal dependence of Superpixels. For a video with a total of T frames, the movement of a superpixel S_i is described by a total of T positions which collectively form the superpixel trajectory, $\mathcal{T}_i$ of superpixel S_i.

$$\mathcal{T}_i := \{(x_0^i, y_0^i), (x_1^i, y_1^i), (x_2^i, y_2^i)...(x_t^i, y_t^i) | t = 1, .., T\} \tag{4.4}$$

We differentiate forward trajectories derived from propagating a superpixel forward from frame $t = 0$ to $t = T$ and backward trajectories from propagating superpixels backwards in time from frames $t = T$ to $t = 0$ using a superscript, $\mathcal{T}_i^F$ or $\mathcal{T}_i^B$ respectively. If a superscript is omitted, forward trajectories are assumed. For the position at time t we use $S_i(x_t, y_t)$ to denote the centroid of superpixel S_i for forward tracks and $S_i(x_{-t}, y_{-t})$ to denote the centroid of superpixel S_i for backward tracks.

Defining the Mesh. In this thesis we use the term 'mesh' to refer to a weighted graph, $G = (V, E, w_v, w_e)$ comprising a set V of vertices, nodes or points together with a set E of edges where w_v, w_e are weight functions that map a vertex, v and an edge e to the set of real numbers, $w_v : V \mapsto \mathbb{R}$ and $w_e : E \mapsto \mathbb{R}$ respectively.

Superpixel Neighbourhood. In order to form a mesh, G from the superpixel positions at time t, we define the (ε-) neighbourhood of superpixel S_i, $\mathcal{N}_\varepsilon^i$ (see also Giraud et al. 2017 where a superpixel and its neighbourhood is termed a 'superpatch').

$$\mathcal{N}_\varepsilon^i = \{S_j \in \mathcal{S} | d(S_i(x_t, y_t), S_j(x_t, y_t)) \leq \varepsilon\} \tag{4.5}$$

where ε is a threshold function and $d(\cdot)$ is a distance function comparing the closeness of two points. A superpixel S_j in the neighbourhood of S_i is connected to S_i by an edge, E_{ij} with an edge 'weight', $w_{e_{ij}}$ and w_{v_i} is a feature vector that describes attributes of the superpixel e.g. the average image intensity at $S_i(x_t, y_t)$.

Dynamic Superpixel Mesh. A dynamic superpixel mesh is the set of all meshes defined by the superpixel neighbourhood from time $t = 0$ to $t = T$.

$$\mathcal{G} := \{G_0, G_1, G_2, ..., G_t | t = 0, 1, ..., T\} \tag{4.6}$$

4.3.2 Specific Superpixel Meshes

We describe the four types of meshes used in this thesis; the MOSES, radial counts density, radial neighbours and kNN mesh, differing in their specification of the vertex weight function w_{v_i}, distance function $d(\cdot)$ and the distance threshold ε. All four meshes are related but seek to highlight quantitatively different features of the motion which may best suit the scientific question of interest. By comparing mesh deformations relative to the initial mesh geometry at each point in time, the MOSES mesh is suited to highlight regions where superpixels are moving respective to their initial neighbours. The radial counts density mesh captures changes in the local density of superpixels over time. It is thus useful to highlight 'sources' and 'sinks' of motion. The radial neighbours mesh achieves a similar effect to the radial counts density mesh but weights the contribution to the superpixel density by scoring neighbouring superpixels based on distance. As such the radial neighbours mesh will be less sensitive in highlighting motion 'sources' and 'sinks' with low superpixel coverage but is better able to highlight motion 'sources' and 'sinks' in regions with

high superpixel coverage more 'confidently' than the radial counts density mesh with greater signal. Finally the kNN mesh extends the distance definition of neighbour in the radial neighbours mesh by using a topological measure of 'distance', the number of nearest neighbours to better describe phenomena where the topology and not the physical distance matters such as bird flocking, (Bialek et al. 2012)

.

The visual differences between the mesh definitions and their dependence on the number of superpixels can be shown by deriving the respective motion saliency maps (Ch.4.4.3) for a few relatable actions from the UCF101 human action recognition dataset (Soomro et al. 2012), (Table.2.3). For completeness we also include comparison to RMSD which does not consider the separation between neighbouring superpixels. The result is shown in Fig.4.6 and discussed in more depth in the respective subsection Ch.4.6.1). Note all the mesh definitions below enforce positivity of w_{v_i}. This is to ensure no distinction between compression and expansion of the mesh during comparison. Additionally it has the benefit of enforcing a degree of temporal ordering c.f. frame embedding (Ch.4.6.1 and Ch.4.6.4).

The MOSES Mesh

The MOSES mesh uses a constant distance cut-off threshold, with a euclidean distance. Critically its vertex weighting function is the change in the average neighbourhood mesh strain with respect to the initial mesh configuration in analogy with mechanics. The superpixel neighbourhood is fixed to be the same as at $t = 0$ for all time.

$$\begin{aligned} d(S_i^t, S_j^t) &= \sqrt{(x_t^i - x_t^j)^2 + (y_t^i - y_t^j)^2} \\ \varepsilon &= cw_s, \quad c \in \mathbb{R} \\ \mathcal{N}_\varepsilon^i(t=t) &= \mathcal{N}_\varepsilon^i(t=0) \\ w_{v_i}(t) &= \mathbb{E}[|d(S_i^t, S_j^t) - d(S_i^0, S_j^0)|], \quad \forall j \in \mathcal{N}_\varepsilon^i(t) \end{aligned} \tag{4.7}$$

MOSES Mesh

Intuitively it can be seen as a local neighbourhood extension of RMSD, the square root of the averaged MSD (Eqn.(3.2)) and related to the Cauchy-Green deformation tensor, a measure of the local material strain rate in continuum mechanics (Eqn.(3.23)).

The Radial Counts Density Mesh

The radial counts density mesh uses a constant distance cut-off threshold with a euclidean distance. Its vertex weighting function is the count of the number of neighbours in the ε-neighbourhood or the degree of a graph. The superpixel neighbourhood is recomputed using the superpixel positions at time t, (Eqn.(4.5)).

$$\begin{aligned} d(S_i^t, S_j^t) &= \sqrt{(x_t^i - x_t^j)^2 + (y_t^i - y_t^j)^2} \\ \varepsilon &= cw_s, \quad c \in \mathbb{R} \\ w_{v_i}(t) &= |\mathcal{N}_\varepsilon^i(t)|, \quad \forall j \in \mathcal{N}_\varepsilon^i(t) \end{aligned} \tag{4.8}$$

Radial Density Mesh

The Radial Neighbours Mesh

The radial neighbours mesh uses a constant distance cut-off threshold with a euclidean distance. Its vertex weighting function is the change in the average neighbourhood mesh strain with respect to the initial configuration where the superpixel neighbourhood is recomputed using the superpixel positions at time t, Eqn.(4.5).

$$\begin{aligned} d(S_i^t, S_j^t) &= \sqrt{(x_t^i - x_t^j)^2 + (y_t^i - y_t^j)^2} \\ \varepsilon &= cw_s, \quad c \in \mathbb{R} \\ w_{v_i}(t) &= \mathbb{E}[|d(S_i^t, S_j^t) - d(S_i^0, S_j^0)|], \quad \forall j \in \mathcal{N}_\varepsilon^i(t) \end{aligned} \tag{4.9}$$

Radial Neighbours Mesh

The kNN Mesh

The kNN (k-nearest neighbours) mesh is a topological distance which defines the neighbourhood of a superpixel with respect to its nearest k superpixels. The euclidean metric is used to measure distance. Its vertex weighting function is

the change in the average neighbourhood mesh strain with respect to the initial configuration where the superpixel neighbourhood is computed using the superpixel positions at time t,

$$
\begin{aligned}
d(S_i^t, S_j^t) &= \sqrt{(x_t^i - x_t^j)^2 + (y_t^i - y_t^j)^2} \\
\varepsilon &= d(S_i^t, S_k^t), \\
S_j &\in \mathcal{S} | d(S_i^t, S_1^t) < d(S_i^t, S_2^t) < ... < d(S_i^t, S_k^t) \\
w_{v_i}(t) &= \mathbb{E}[|d(S_i^t, S_j^t) - d(S_i^0, S_j^0)|], \quad \forall j \in \mathcal{N}_\varepsilon^i(t)
\end{aligned} \tag{4.10}
$$

kNN Mesh

4.4 General Motion Analysis Tools

The analysis of biological motion due to the wide diversity and inherent variation of the specific biological processes being studied makes it challenging to provide a single method to analyse the generated tracks. Tailored approaches must often be developed to answer specific hypotheses based on domain specific knowledge. However, certain components of the downstream process can be generalised and below we propose three general methods, two based on the developed notion of dynamic meshes to facilitate visualization of video motion and exploratory analysis suitable for any video.

4.4.1 Motion Track Clustering

When dealing with a large number of superpixel motion tracks that are spaced closely together, it is often useful to group them together spatially and with similar trajectories for interpretation. This can be done through clustering. There are numerous approaches but we found that applying GMM (Gaussian Mixture Model) clustering after dimensional reduction of the tracks was effective, offering a good compromise between accuracy and speed. The procedure is as follows (c.f. Zhang et al. 2006).

For each superpixel trajectory $\mathcal{T}_i$ from frame 0 to the desired frame T, form the concatenated 1d signal of all (x, y) coordinate positions over the entire time interval denoted $\mathcal{T}_i^{\text{Flattened}}$.

$$\mathcal{T}_i^{\text{Flattened}} = [x_0^i, x_1^i, ..., x_T^i, y_0^i, y_1^i, ..., y_T^i] \tag{4.11}$$

If any trajectory is not present over the whole time interval $[0, T]$, padding and interpolation is implemented to ensure all trajectories for clustering are of the same sequence length. The trajectory replicates the first recorded position up to time $t = 0$ and replicates the last recorded position up to time $t = T$. In between, linear interpolation is used. For N superpixels then stack the flattened trajectories into a $N \times 2T$ matrix and apply PCA choosing the number of components, K to be significantly smaller than $2T$, $K << 2T$ e.g. 5 or 10 for $T = 100$. This yields a new $N \times K$ matrix which can be clustered using GMM with a user specified number of components. The higher the number of GMM components, the smaller the resulting group size.

4.4.2 Motion Signature Generation

It is difficult to directly use dynamic meshes to compare across videos as the dynamic mesh in general requires a $(T \times N \times N_{features})$ matrix to describe the vertex features and a $(T \times N \times N)$ matrix to describe the dynamic edge connections for a video with T frames, N superpixels and each vertex is described by $N_{features}$ features. For the MOSES mesh with $N_{features} = 1$ we illustrate how to derive three motion signatures and one spatial signature that can be used for similarity comparison of videos within a video collection and video frames within a single video.

1.Forward Motion Signature

$$\text{Sig}_{\text{MOSES}}^F(t) = \left\{ \frac{1}{|\mathcal{S}|} \sum_i w_{v_i}(t) | t = 0, 1, ..., T \right\} \tag{4.12}$$

where $|\mathcal{S}|$ is the total number of superpixels, the cardinality of the superpixel set. The forward motion signature is also referred to as the MOSES mesh strain curve. Theoretically, it captures where motion flows to c.f. forward FTLE (Ch.3.2.2).

The normalised MOSES mesh strain curve divides the MOSES mesh strain curve by its maximum value.

2. Backward Motion Signature

$$\text{Sig}^{B}_{\text{MOSES}}(t) = \left\{ \frac{1}{|\mathcal{S}|} \sum_{i} w_{v_i}(t) | t = T, T-1, ..., 0 \right\} \tag{4.13}$$

where $|\mathcal{S}|$ is the total number of superpixels, the cardinality of the superpixel set. Theoretically this captures where motion flows from c.f. backward FTLE (Ch. 3.2.2).

3. Total Motion Signature

$$\text{Sig}^{T}_{\text{MOSES}}(t) = \frac{1}{2}\Big(\text{Signature}^{F}_{\text{MOSES}}(t) + \text{Signature}^{B}_{\text{MOSES}}(t)\Big) \tag{4.14}$$

Together this captures the motion boundaries of long-term flow c.f. LCS (Ch. 3.2.2).

Spatial Motion Signature for individual video frames For a single video with the MOSES mesh where the edge connections remains the same for all time, and the total number of features is the same, the $(T \times N)$ features matrix can be viewed as a spatial signature with each of the T frames defined by the respective N vertex weights, w_{v_i}.

$$\text{Sig}^{Frame}_{\text{MOSES}}(i, t) = \{ w_{v_i}(t) | i \in \mathcal{S} \} \tag{4.15}$$

4.4.3 Motion Saliency Map

The core of video analysis is to identify the primary moving objects of interest and measure their characteristics as they move in time. Ideally one could segment the subjects of interest from the background but often this is very difficult, (Ch.2.3.3). What if coarse localisation of the object(s) of interest with a bounding box was sufficient? such as in automatic camera following of a particular cell maintaining it within the centre of the field of view or the detection of unusual spatial regions of motion. For such questions, deriving a motion saliency spatial heatmap can be used as a powerful visualisation technique and to provide a quick exploratory solution

that aids in downstream applications. The concept of visual saliency has been well explored before in the literature e.g. for images (Itti et al. 1998; Harel et al. 2007) and for video (Itti and Baldi 2005; Fragkiadaki et al. 2012). In Itti and Baldi 2005, saliency is defined as a local 'surprising' event where a pixel's image features are different to its local neighbours and elicit more motion activity. Centre surround filtering based on various cues is proposed to uncover the motion saliency. Our approach is most similar to the trajectory embedding motion saliency of Fragkiadaki et al. 2012 but simpler. No embedding is required and the motion saliency can be computed frame-to-frame or for a subset of frames.

With dynamic meshes we can use non-grid-like graphs to extend the notion of the Finite Time Lypanunov Exponents (FTLE) field. With forward trajectories this leads to discovering 'sinks' of motion, where does motion move towards? whilst backward trajectories leads to identification of 'sources' of motion, where do motion originate from? We refer to the resulting map as a motion saliency map, $Sal(x, y)$ which we define in this thesis as a temporally averaged image across a time interval $[t_0, t_0 + \delta t]$. For each time frame a spatial image is formed the same size as the video frame where the vertex weights w_{v_i} of each superpixel is mapped to the corresponding superpixel centroid position in the image at time t, (x_t^i, y_t^i). If two superpixel areas overlap, their corresponding vertex weight is added, 'upvoting' this spatial area accordingly:

$$Sal(x, y, t) = f(x, y) * \frac{1}{\delta t} \left\{ \sum_{t=t_0}^{t_0+\delta t} \mathbb{P}\left[w_{v_i}(x_t^i, y_t^i) | i \in \mathcal{S} \right] \right\} \tag{4.16}$$

Motion Saliency Map

$f(x, y)$ is a 2D smoothing kernel typically a 2D Gaussian, $f(x, y; \sigma) = \frac{1}{2\pi\sigma^2} e^{-\frac{x^2+y^2}{2\sigma^2}}$, $*$ is a convolution operation and $\mathbb{P}[\cdot]$ is an estimated spatial probability distribution of the weights here using a 2d histogram where the image is equipartitioned into K regular square bins. For bin k with width w, $[x_k < x < x_k + w, y_k < y < y_k + w]$ we assign the total sum of w_{v_i} of all superpixels whose centroids lies in it.

$$\mathbb{P}_k = \sum_i w_{v_i}^t, \quad \{\forall i \in \mathcal{S} | x_k < x_i^t < x_k + w, y_k < y_i^t < y_k + w\} \tag{4.17}$$

This probability distribution estimation is used to 'densify' the superpixel points whose number is typically much smaller than the number of image pixels to create a smoothly varying heatmap. Unless stated otherwise, in this thesis the video motion saliency map is usually computed over the entire length of the video, $t_0 = 0$ and $\delta t = T$ to locate temporally stable motion 'sources' and 'sinks'. In Ch.6.5.4 we demonstrate how one can compute and use the motion saliency map from frame-to-frame, in this instance for organoid branch tracking.

4.5 Experimental Validation of Superpixel Tracking

Whilst the notion of tracking superpixels as image atoms for uncovering the motion patterns is widely used in computer vision with 'natural' images (Brox and Malik 2010; Wu et al. 2011; Wang and Schmid 2013; Wang et al. 2013), in biology it is important to also be able to interpret the biological relevance of the superpixels and to attribute their movement to single or groups of cells. Surprisingly, despite the widespread use of Particle Image Velocimetry (PIV), equivalent to the tracking of fixed square superpixels for studying cellular motion in confluent monolayers when accurate individual cell segmentation is not possible, quantitative association with single cell tracking has not really been addressed. The use of PIV has primarily been justified indirectly by comparing derived measurements, such as speed (Petitjean et al. 2010), or by analogy of the epithelial sheet motion to fluid-like dynamics (Szabo et al. 2006a; Angelini et al. 2010; Angelini et al. 2011), or reasoned from force measurements (Trepat et al. 2009) or qualitative comparison (Zaritsky et al. 2012b). To test the accuracy of superpixel motion extraction directly in the setting of collective cell migration, experimentally a sparse third population of blue cells was added to one of two epithelial sheets separated by a defined gap size (the blue cells being the same cell type as the population to which they were added), (Fig.4.2). For generality, different combinations of cell types (EPC2:EPC2, EPC2:CP-A and EPC2:OE33) of three different immortalised cell lines EPC2, CP-A and OE33 was used. Further the red/green dye used to label cells and the spiked-in population

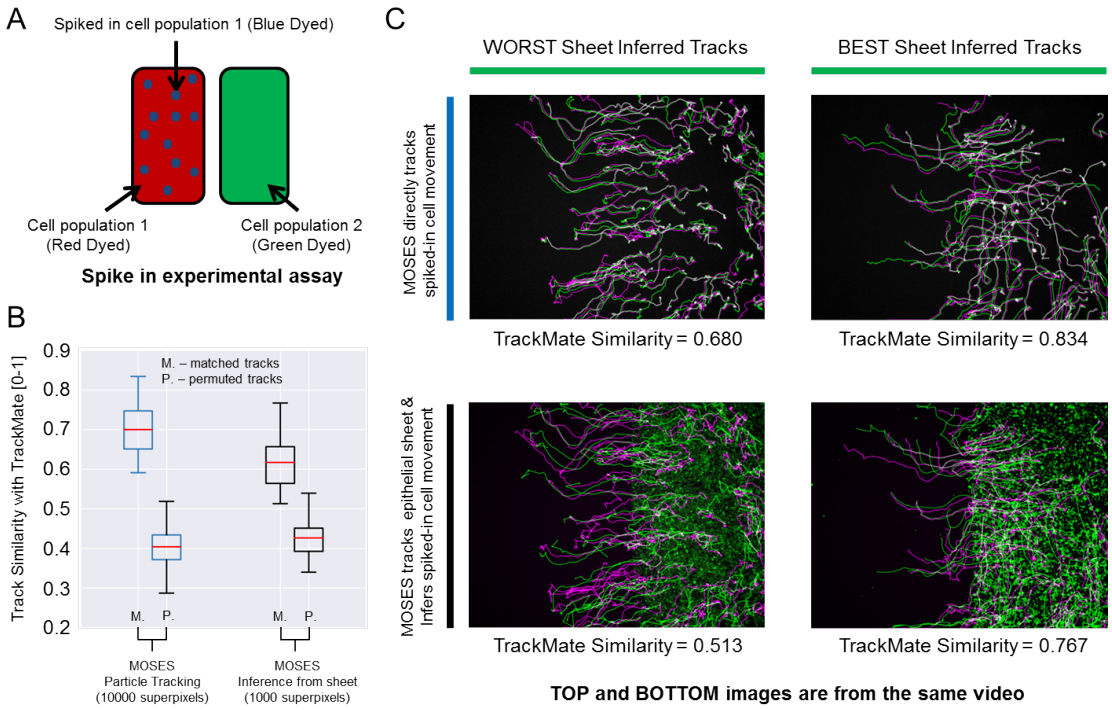

Figure 4.2: Experimental validation of MOSES optical flow superpixel tracking. (A) Schematic illustration of spike-in experimental setup. A separate sample of cell population 1 or 2 is dyed a third colour (blue; dots on figure) and added to the respective red or green sheet (solid colours on figure). (B) Comparison of MOSES tracks of spiked-in cells and the benchmark single cell TrackMate tracks when the blue spiked-in cells were explicitly tracked (blue boxplots, 10000 superpixels) or indirectly inferred from the coloured sheet (black boxplots, 1000 superpixels). Red centre line=mean. Track similarity is reported using normalised track cross-correlation after matching the TrackMate and MOSES tracks (M) and if the correct matchings were randomly permuted (P). Permuted track similarity values for each video are the average of 10 random permutations. (C) Track results of the best and worst inferred tracks from sheet motion according to track similarity and the respective single cell spiked-in MOSES tracks overlaid on a video snapshot at 0 hr. Magenta: TrackMate tracks. Green: MOSES tracks of green cell population. Tracks are white if TrackMate and MOSES perfectly coincide.

were alternated, generating 23 videos (each 144hr). To assess the performance of MOSES in tracking with relation to particle level resolution (i.e. tracking the individual spiked-in cells), tracks generated by TrackMate (Tinevez et al. 2017), a readily available published single particle tracker implemented in Fiji (Schindelin et al. 2012) was used as a benchmark for comparison. Any other suitable single cell or single particle tracking could have been used as benchmark such as U-track Jaqaman et al. 2008. The exact algorithm used was not important as we are not comparing the ability to handle new cells or cell divisions, see below and all

MOSES Parameters	MOSES Values	TrackMate Parameters	TrackMate Values
number of superpixels	1000 or 10000	est. blob diameter	10 pixel
est.blob diameter	multiscale, (1-10 pixel at 5 equally spaced intervals)	image intensity threshold	2.5
threshold	0.05	filter on spots uniform colour, LAP tracker	
		frame-to-frame linking max distance	50
		track segment gap-closing max distance	50
		track segment gap-closing max frame gap	100

Table 4.1: Table summarising the MOSES parameters and the TrackMate parameters used for validating superpixel tracking.

produced single-cell tracks were manually checked.

To assign tracks to the blue spiked-in cells with MOSES (Step 4, Fig.4.1), a multiscale (5 scales, c.f. Table.4.1) LoG (Laplacian of Gaussian) blob detector was used to segment the cell area. For each segmented cell, of the superpixels that lie within the area of the cell, the longest track was assigned to describe the motion of the cell. To generate single cell tracks from the sheet tracks using MOSES, for each segmented cell, the nearest 4 superpixel tracks from the respective epithelial sheet (red or green) were found to compute a mean track to estimate its motion. The benchmark single cell tracks was obtained by running the Fiji TrackMate plugin on only the image channel containing the sparse population. The computational parameters used for each method is summarised in Table.4.1.

Track similarity (discussed in Ch.3.3.4) is given by the normalised track cross-

correlation (value between 0,1) between each MOSES track and the corresponding TrackMate track. The average track similarity is reported for each video (Fig.4.2B). Formally for a track i and track j, the track or trajectory cross-correlation is defined as:

$$\begin{aligned} \mathrm{TCC}_{ij}(m,t) &= \left\{ \frac{1}{T} \sum_{n=-T}^{T} \hat{\mathcal{T}}_i(t+m) \cdot \hat{\mathcal{T}}_j(t) \right\} \\ \mathrm{TCC}(\mathcal{T}_i, \mathcal{T}_j) &= \max\left(0, \mathrm{TCC}_{ij}(m_{\max}, t)\right) \\ \text{where} \quad m_{\max} &= \arg\max_{m} \{\mathrm{abs}(\mathrm{TCC}_{ij}(m,t)) | m = -T, ..., T\} \end{aligned} \tag{4.18}$$

where abs$(\cdot)$ takes the absolute value, $\hat{\mathcal{T}}_i = \frac{\mathcal{T}_i - \bar{\mathcal{T}}_i}{\sigma_i}$ and σ_i is the standard deviation of $\mathcal{T}_i$. T is the maximum time lag and is the shorter of $\mathcal{T}_i$ or $\mathcal{T}_j$, $\min\{L_{\mathcal{T}_i}, L_{\mathcal{T}_j}\}$, t is time and m the time lag. This definition makes no distinction between negative or positive correlation values. This may be suitable for applications where identifying the presence or absence of correlation is sufficient e.g the interaction between two epithelial sheets in Ch.5.2.3. Where it is desired to compare how identical two tracks are we modify Eqn.(4.18) such that negatively correlated tracks returns a value of 0.

To assess the statistical significance of the resultant correlation value, the track similarity from random pairings of the tracks were computed and the average of 10 permutations were reported (labelled P. for permuted in Fig.4.2B) per video. Since MOSES does not explicitly handle cell division or the introduction of new cells that emerge in the field of view, to ensure fair comparison only the tracks belonging to the initial cells are compared. When the field of view was divided into 10000 superpixels in MOSES, the tracks were highly similar to TrackMate, (Fig.4.2B) and by visual assessment of the worst and best inferred tracks, (Fig.4.2C top). To assess if the motion of single cells could be inferred from sheet motion, 1000 superpixels were used to track the sheet motion and the estimated single cell trajectories compared to TrackMate. Although the precise individual cell trajectory was lost, the overall motion pattern of the spiked-in cells could be inferred (Fig.4.2B,C bottom). Therefore, in confluent epithelial sheets, individual cells behave similarly

to their neighbours confirming that global motion patterns can be used as a proxy to study single cell behaviour. Each superpixel (chosen sufficiently small) can represent the dynamics of a local cell group.

4.6 Applications

As further validation of MOSES and proof of its general applicability and utility for exploratory analysis of general video datasets, MOSES was used to analyse four different video datasets, ranging from very distinct areas of biology to a general video dataset UCF101 (Soomro et al. 2012), an extensively used action recognition benchmark in the computer vision community comprised of clips from YouTube. We use the UCF101 dataset as its motions are easily relatable and there are manually annotated labels. It is ideal to explore the core principles and ideas of MOSES; its advantages and limitations.

4.6.1 Motion Saliency in YouTube Videos

An important video classification problem is action recognition, the ability to categorise a short video snippet according to the salient action present (Ch.2.3, 2.3.1, 2.3.2). Several benchmark datasets have been published, (Table.2.3). UCF101 is one such dataset which aims to capture 'realistic' action videos. It comprises 13320 videos from YouTube organised into 101 distinct action categories such as golfing, rowing, playing piano and further grouped into 5 superclasses according to motion type 1) human-object interaction, 2) body-motion only, 3) human-human interaction, 4) playing musical instruments and 5) sports, (Fig.4.3). With a diverse set of different actions, the additional introduction of noise due to background movement, camera shake, camera zoom and the presence of multiple moving subjects, UCF101 is an ideal dataset to test the proposed motion signatures and motion saliency maps. With this in mind, three experiments were carried out to assess how specific and generally applicable MOSES is across 'datasets' that capture different distinct motion patterns where we can think of each action class as a distinct 'dataset'. In analogy with cellular motion, the action classes could represent

separate drugs. It should be noted that the ensuing experiments using UCF101 was largely qualitatively assessed. Only top-level action classification labels are provided. There is no ground-truth motion saliency maps or finer temporal segmentations of the individual actions. The purpose was not to reach or eclipse the state-of-the-art with MOSES but only to be able to sanity check our proposed definitions of mesh and motion saliency on a well tested video dataset that is relatable to non-experts and easy to work with computationally; one single video clip, one salient action, small enough to download and distribute and easy to set up.

Experiment 1: Can MOSES Capture Salient Motion in the presence of Background Motion?

The ability to summarise motion patterns in arbitrary length videos as a single static image for visualization is extremely useful for exploratory data analysis. The ability of the proposed MOSES motion saliency map (Ch.4.4.3, Eqn.(Motion Saliency Map)) defined on dynamic meshes to localise the most salient motion patterns spatially was tested using UCF-101. 1000 superpixels was used throughout. The 'raw' tracks from motion extraction (Steps 1-3, Fig.4.1) was used without further filtering. Visually we compared to state-of-the-art dynamic images, (Bilen et al. 2016a; Bilen et al. 2016b), another static image summary. The mean and maximum video image formed by taking the mean and maximum pixel intensity values across time was also computed to give a summary of the range of motions present in each video. A dynamic image is a compact representation of a video as a single static image by training a ranking machine to learn the temporal ordering of frames. This idea is known as rank pooling. The central premise is that good spatiotemporal features are learnt if a classifier is able to accurately recover the temporal order when fed the video frames in a random order. A consequence of this ranking operation as noted by the authors is the propensity for the resulting dynamic images to focus on the identity and motion of the salient actors in the videos whilst averaging out background pixels and background motion patterns along with an ability to distinguish between actions of different speeds, (Bilen et al. 2016b). This ability to capture salient motion with dynamic images was demonstrated by reporting (at the

Figure 4.3: The UCF101 action recognition dataset. Still frame snapshots illustrating each of the 101 curated action classes from YouTube taken from the website, http://crcv.ucf.edu/data/UCF101.php. The box colours corresponds to the five supertypes the 101 actions were divided into, 1) human-object interaction 2) body-motion Only 3) human-human interaction 4) Playing Musical Instruments and 5) sports.

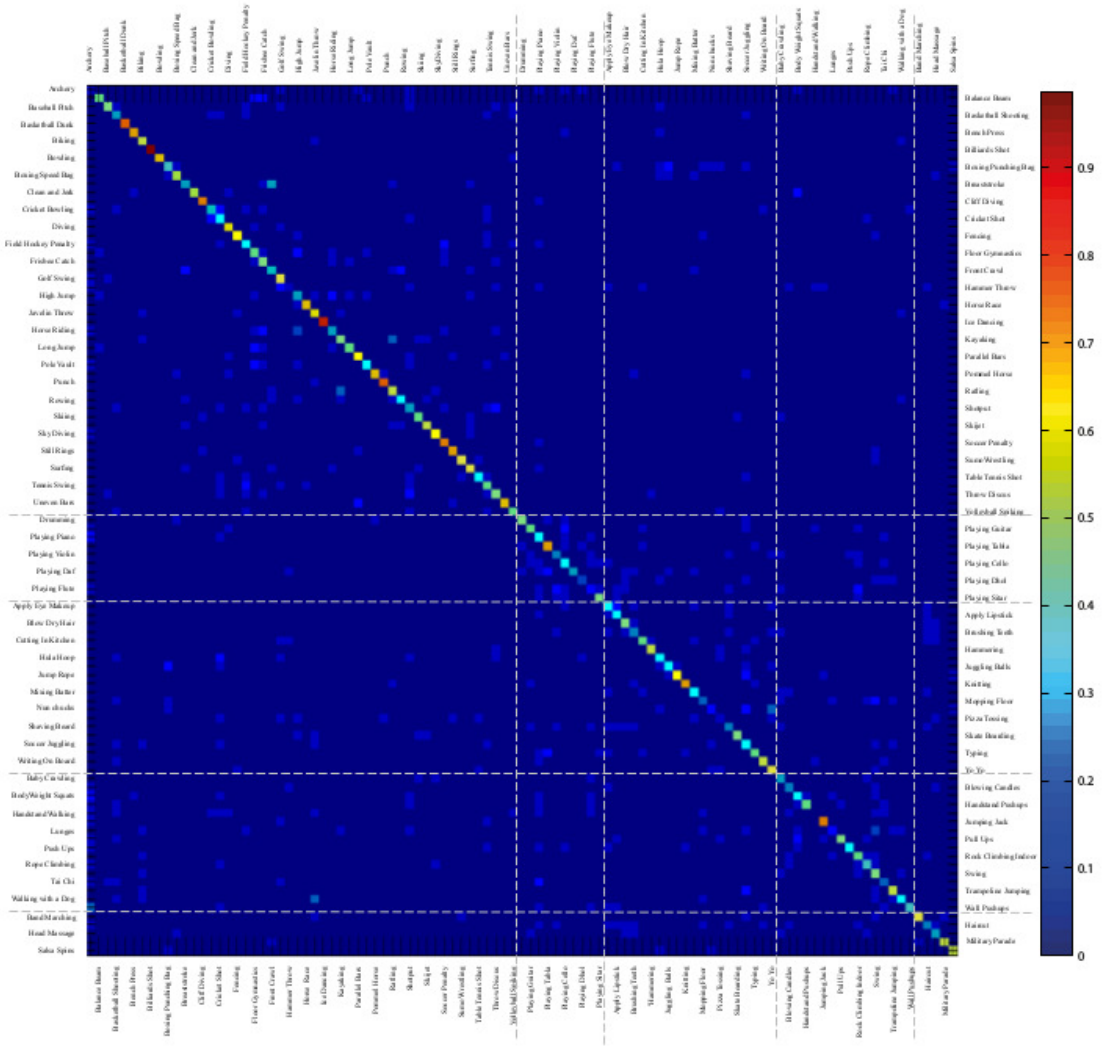

Figure 4.4: Confusion table of baseline action recognition results on UCF101 reported in Soomro et al. 2012. A score of 1 indicates perfect recognition and 0 complete misclassification. The drawn lines separate the different super-types of actions in Fig.4.3; 1-50: Sports, 51-60: Playing Musical Instruments, 61-80: human-object interaction, 81-96: body-motion Only, 97-101: human-human interaction.

time) state-of-the-art classification results on UCF101 and HMDB51.

The results are shown for 10 clips in Fig.4.5 for 10 different actions, 2 each from the 5 super-types, one that is poorly classified and one that is well classified according to the baseline confusion matrix in the original publication of Soomro et al. 2012; archery, playing piano, pizza tossing, handstand walking and haircut are poorly classified, ice dancing, playing tabla, knitting, jumping jack and band

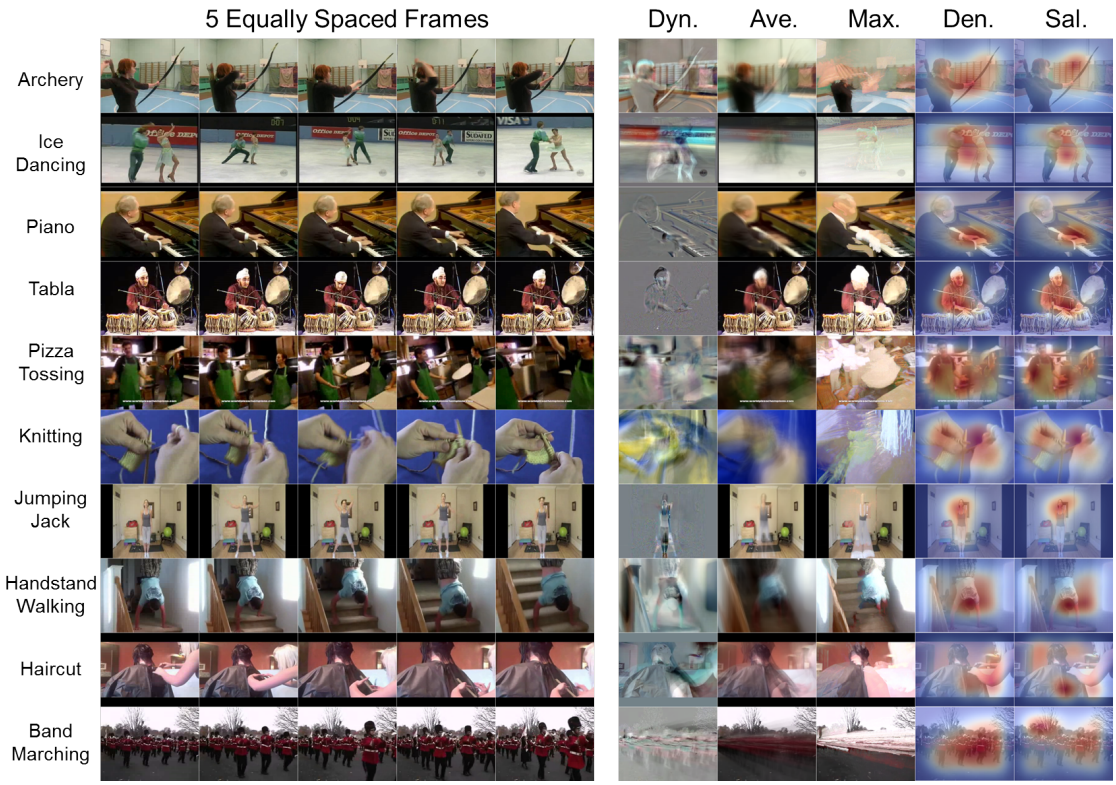

Figure 4.5: UCF101 motion saliency maps for 10 action classes. Videos are of different lengths. 5 equally spaced static snapshots are shown. Dyn.-Dynamic RGB image of the video (Bilen et al. 2016a). Ave.-Average or mean image of the video stack. Max.-maximum projected image of the video stack. Den.-Density heatmap image by considering all the superpixel trajectories as points and plotting them all onto the same image. Sal.-MOSES saliency map, (Ch.4.3.2, Eqn.(MOSES Mesh), Ch.4.4.3, Eqn.(Motion Saliency Map)).

marching are well classified. Together they represent a mixture of subtle and large-scale motion from 'everyday' scenes. For each video, 5 equally spaced frames is visualised to illustrate the motion dynamics. Approximate rank pooling (Bilen et al. 2016a) was used to compute the dynamic image as it is very efficient with little loss in performance (96.5%±0.9 vs 99.5%±0.1 temporal ordering accuracy). As seen in Fig.4.5 Dyn., Dynamic images provide an excellent dynamic summary when the motion is large and distinct such as jumping jack or band marching and is able to blur out the background in these actions using grey pixels, is confusing or difficult to interpret with complex motion such as pizza tossing or knitting and cannot capture convincingly localised complex dynamics within only a small spatial area such as the hands of the piano player or the snipping of the hairdresser. In comparison the average image (Ave.) excessively blurs out fast motion such as the ice dancing or

pizza tossing and knitting but is able to identify regions of subtle motion such as the piano or tabla. Finally the maximum (Max.) image excels at exacerbating subtle motions, excessively highlighting the drawing action of the archer but is difficult to interpret. Thus while dynamic, average and maximum images can help indicate the direction of motion, it is poor at spatially localising the motion. The point density of the superpixel tracks over time (Den.), counting the number of times a superpixel is at a particular (x, y) position in the image, i.e. $w_{v_i} = 1$ is a good baseline in localising motion highlighting all image pixels experiencing more motion than their surroundings, (Fig.4.5 Den.). However this detection is non-specific and does not find the most important essence of the movement that may help classify the action. In the playing of the Tabla, Den. highlights the body as the most salient whereas it is the hands that is critical to the action. The MOSES motion saliency map, (Ch.4.3.2, Eqn.(MOSES Mesh)) by utilising neighbourhood information better localises the salient motion. For example it is the only method to correctly identify the snipping action of the hairdresser at the bottom of the image despite the larger brushing action.

Using the same videos, we can assess the impact of the mesh definition and the impact of the number of superpixels for identifying motion saliency. To assess the impact of the mesh definition, the MOSES mesh saliency motion map is compared to the RMSD saliency map, the non-collective equivalent of MOSES and 3 alternative meshes; the radial density mesh, the radial neighbours mesh and the kNN mesh as defined in Ch.4.3.2, (Fig.4.6). With 100 superpixels there is an insufficient number of superpixels to capture the video motion. The resulting motion saliency maps is not specific for any of the 10 videos. For 1000 superpixels, specific motion saliency is captured and with increasing numbers of superpixels e.g. 5000 captures ever finer motion. The motion saliency map is increasingly less smooth with the appearance of more heatspots. It is clear between RMSD and mesh based saliency maps, that the mesh based saliency maps are more robust and tolerant to noise. For example, for handstand walking 1000 and 5000 superpixels, RMSD overemphasises

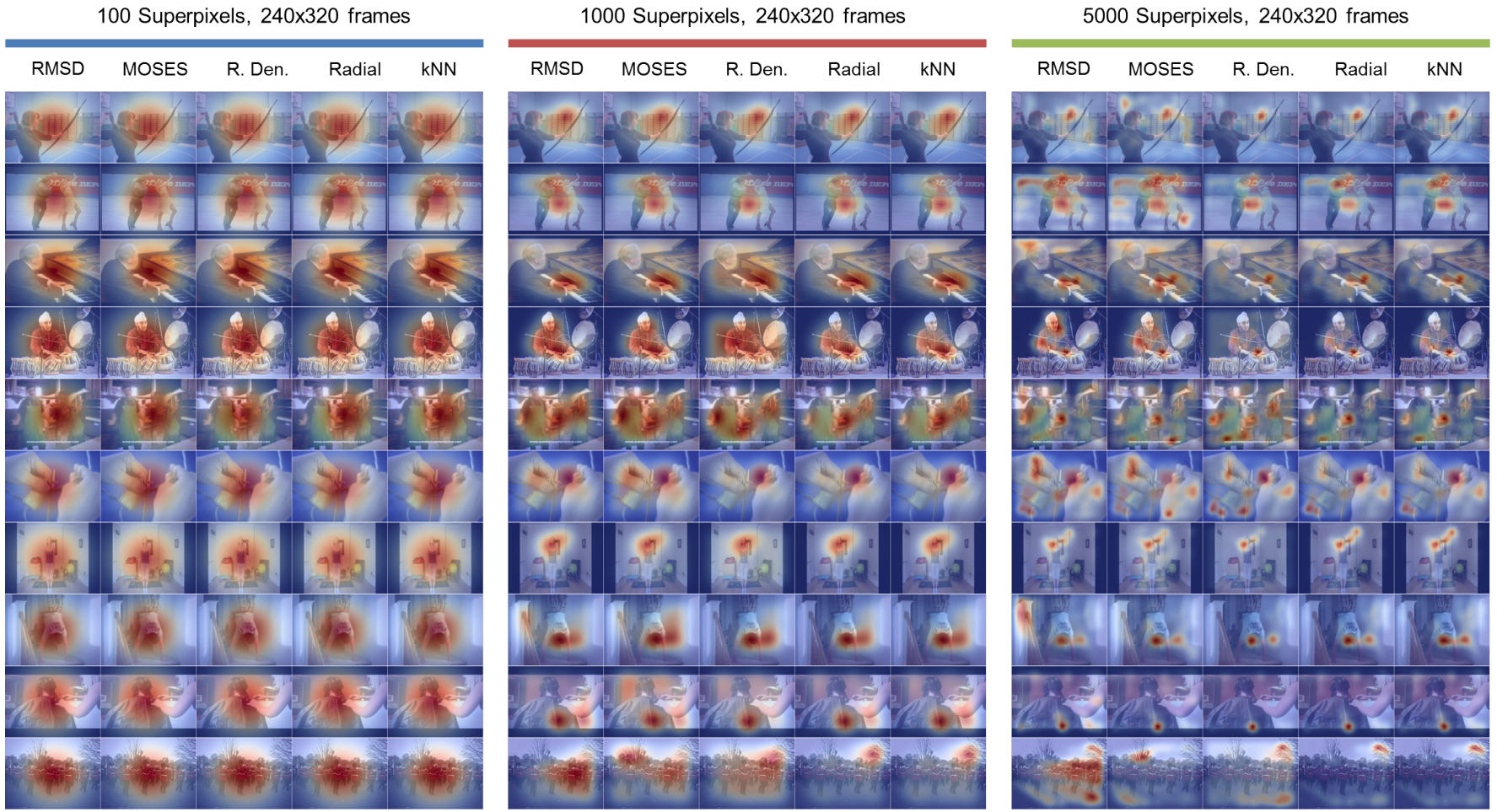

Figure 4.6: The effect of number of superpixels and mesh on the resulting motion saliency maps. The same videos are used as in Fig.4.5. R. Den.- radial density mesh, (Ch.4.3.2, Eqn.(Radial Density Mesh)). Radial - radial neighbours mesh, (Ch.4.3.2, Eqn.(Radial Neighbours Mesh)). kNN - kNN neighbours mesh, (Ch.4.3.2, Eqn.(kNN Mesh)).

the apparent motion of the stairs (left-most heat spot) and similarly for the tabla playing with 5000 superpixels RMSD overemphasises the head movement of the player whilst the arms and hands remain the most important for MOSES. However there is no clear winner amongst the mesh-based definitions. Between MOSES and other mesh variants, there is different pros and cons. For example for ice dancing a long-time, large dynamic motion with 5000 superpixels we get an idea of the spinning nature of the movement, similarly with archery and jumping jack but other variants perform more robustly for subtle motion such as radial density (R. Den.) for piano and tabla playing and radial neighbours (Radial) for complex short-time motion like the hands in pizza tossing.

Experiment 2: Can MOSES Measure Motion Similarity?

A motion signature should be able to compare and retrieve similar videos given a video dataset. As such we tested the capacity of the MOSES mesh to represent uniquely the video motion content in two ways. First we show that given a single video MOSES allows embedding of each frame in a manner that obeys both the motion pattern and the inherent temporal ordering using simple principal components analysis (PCA) on the proposed frame signature, (Eqn.(4.15)). The results for the archery, ice dancing and piano playing is shown in Fig.4.7. Interestingly, the shape of the 2D embeddings appear to reflect the complexity and speed of the motion. In particular, motions with slow appearance changes are grouped together (high density point region with lots of neighbouring embedded frames) and motions with ever changing appearance are spread out (low density point region with few regularly spaced neighbouring embedded frames) whilst respecting the temporal order. For example, in archery g01, the drawing preparatory motions is a continuously changing motion and is drawn out (frames $\approx$1-15) whilst the firing frames is similar and compact. Two shots were fired in this video coinciding with the two dips of increased tortuosity and high point density in the embedding. Similarly, in ice dancing g01, the fast twirls (sparse regions) contrast with the smoothly varying slow skating (dense regions). For piano playing the drawn out regions highlight frames of large

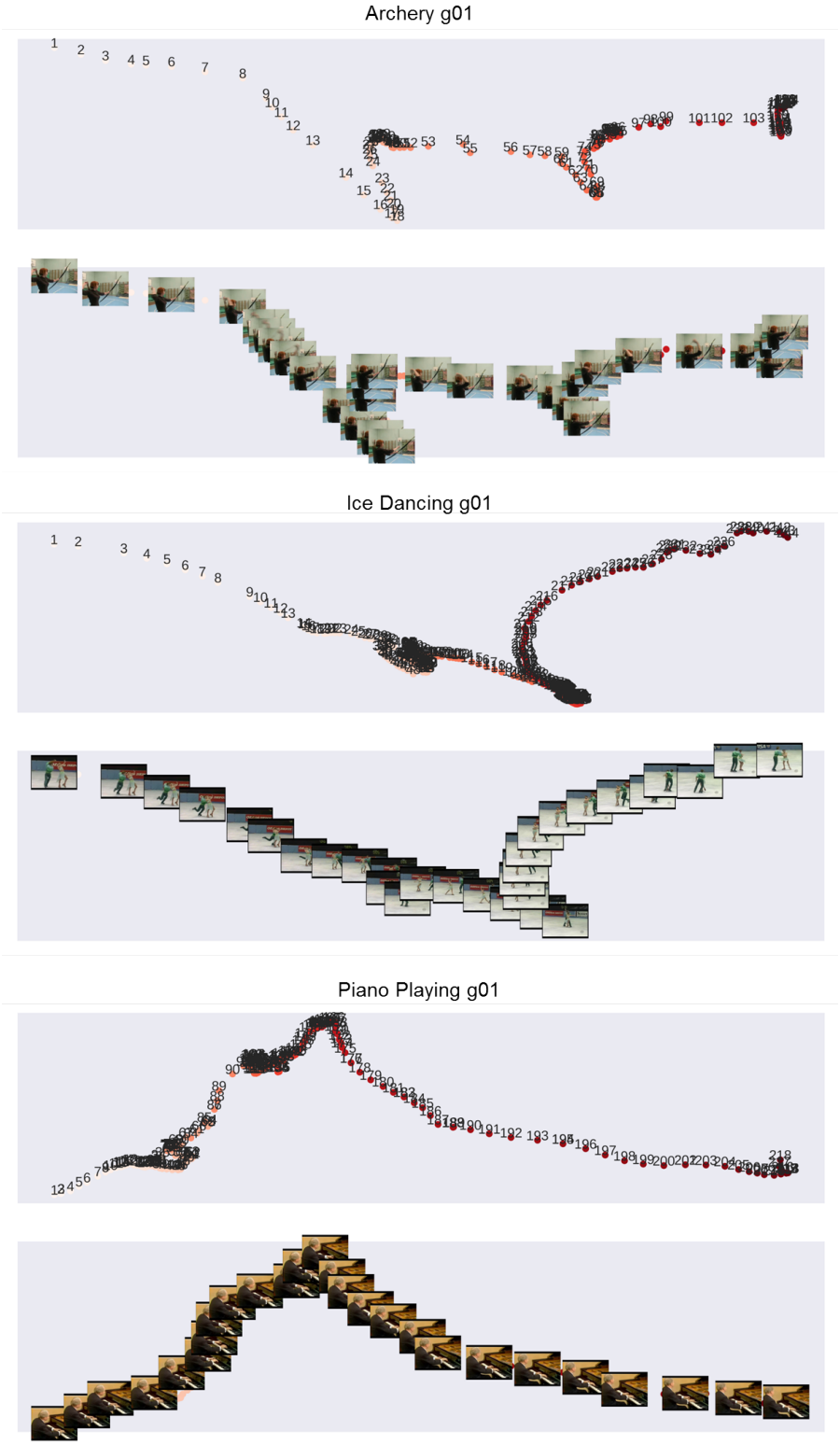

Figure 4.7: Temporal embedding of video frames using PCA on MOSES spatial signatures. 3 videos are shown, PCA was applied to project the frame signature, Eqn.(4.15) into 2D for plotting. For each video, in the upper panel each point is a video frame coloured in increasing red and labelled with increasing time chronologically.

finger movement. Thus the MOSES frame signature naturally respects temporal ordering and distinguishes between slow and fast motions. This observation could be exploited as part of an automatic algorithm to zone in and identify distinctive motion frames within a single video. One such implementation would be to compute the number of neighbours within a certain radius for each video frame following embedding with PCA. A discontinuous jump would be expected when the number of neighbours within a preset radius based on the embedding is plotted against the frame number. Thresholding can then be applied to identify the transition between motion regimes marked by frames with a discontinous transition from low density neighbour frames to high density neighbour frames. A biological application of this concept is to identify the distinct motion regimes in a long time-lapse acquisition such as *Drosophila* embryo development (4.6.4) for automatic video trimming to allow subsequent application of single cell tracking. Doing so allows single-cell tracking to operate on shorter video segments which is more accurate and more computationally efficient.

Second, we took the full set of archery videos, 125 clips from 25 distinct videos and tested if the MOSES forward signatures, (Eqn.(4.12)) had sufficient representation power to facilitate unbiased clustering of videos based on motion alone, (Fig.4.8A left, clips belonging to the same video are marked with the same colour.). This is useful for example to assess the heterogeneity of motion for the same action. In biological applications, it is useful to assess the inherent heterogeneity present in the 'wild-type' condition. Since the videoclips were of different lengths, shortest 51 and longest 557, all MOSES signatures were stretched to the maximum length using spline interpolation. PCA (2-components) was then applied with whitening. The majority of videos mapped to a single dense region, clustering tightly. This suggests MOSES is capturing the action, recognising that most of the videos represent the same action, archery in spite of the confounding factors of different camera angles, colour, frequency of shooting and style. To assess the mapping in further detail, Rasterfairy (Klingemann 2017) a method to replot a 2D set of points as a

regular grid with no overlap whilst preserving their neighbours was used, (Fig.4.8A right, clips belonging to the same video are marked with the same colour.). In the majority of cases, most of the clips from the same video were grouped together. In those that do not MOSES demonstrates the ability to identify 'outlier' videoclips. Here an 'outlier' is understood to be a single videoclip whose nearest neighbours following PCA fails to group with the majority of the other videoclips from the same video. For example for video 1 (g01, highlighted with blue box in Fig.4.8B,C) using Fig.4.8C which pictures the first frame of each clip, one might expect at most 3 groups, ({1,7}, {2,3,4}, {5,6}) but the clips are fragmented according to PCA into 5, (Fig.4.8A right). Analysing the corresponding signature (Fig.4.8B top), we see that our prior prediction based on the single initial frame is incorrect. In reality, 5,6 do not cluster because whilst similar, the action of 6 is more delayed, 2 does not cluster as it should because near the end of the clip, the camera suddenly zooms in and 1,7 does not cluster because there is an additional significant camera pan at the start of the clip. It should be noted action recognition motion signatures should not be affected by camera motion. One way to improve the current analysis with respect to camera motion would be to first estimate and remove the dominant camera motion as in Wang and Schmid 2013 and Jain et al. 2016b prior to computing MOSES signatures. In a second case study, one clip of video 8 (highlighted with red box) is a significant 'outlier' to the other clips from the same video despite having very similar initial appearance, (Fig.4.8B). In this case, however the difference is due to an artefact of the time stretching implemented to ensure the same length signatures. The particular clip was significantly shorter, (Fig.4.8B bottom) but possesses very similar action profiles. To rectify these temporal sampling artifacts, one could trim the signature to match the length of the shortest clip before applying PCA. An alternative signature that requires no resampling (given the same image size) would be to take the signature from averaging over time as opposed to averaging over superpixels (c.f. Eqn.(4.12)) or one could defer handling of the different time lengths to the choice of dimensional reduction or classification algorithms.

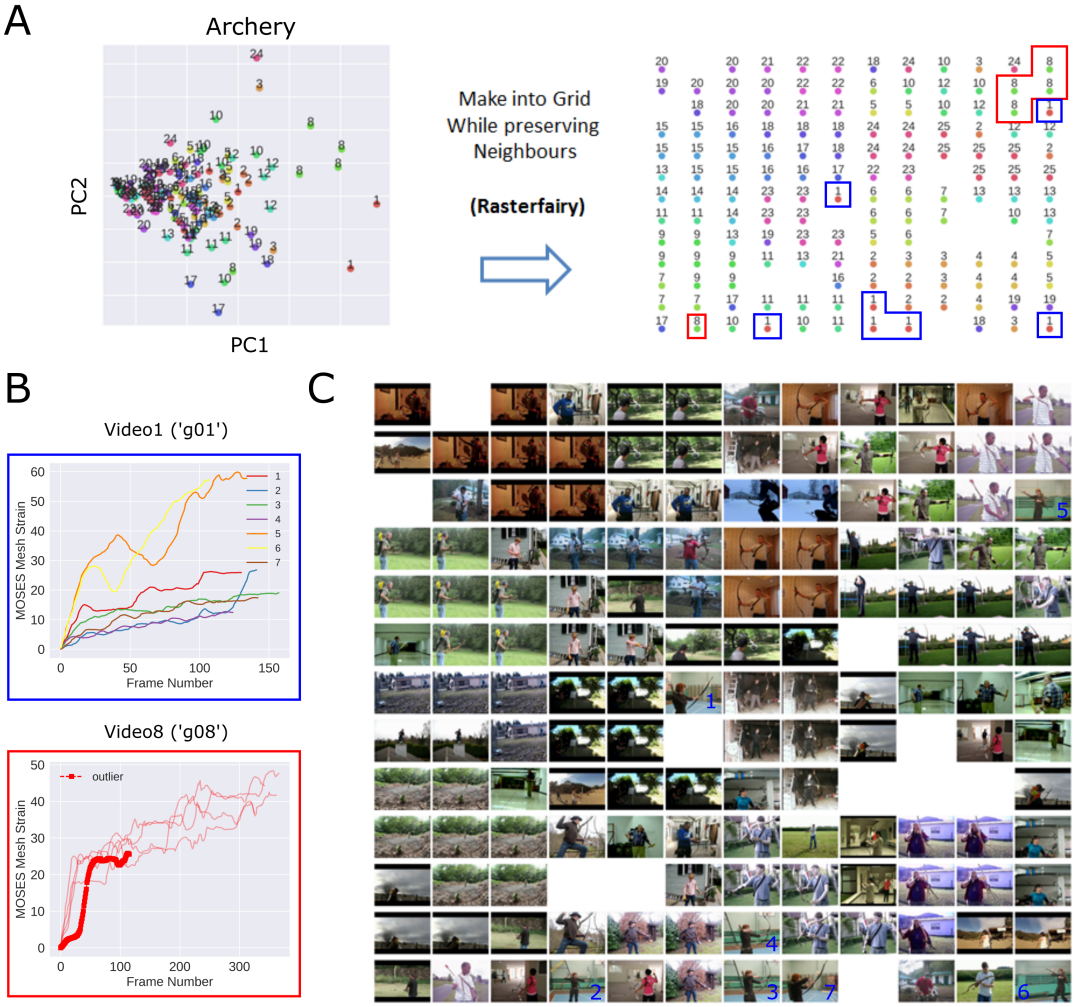

Figure 4.8: Left: MOSES signature for similarity comparison of UCF101 archery videos. A) PCA embedding of all UCF101 archery videos based on the forward MOSES signature, (Eqn.(4.12)). Each point is a video clip, n=125 clips taken from 25 longer videos. Points are coloured the same if they are from the same longer video. Right: remapping of the PCA coordinates into a regular grid with Rasterfairy. B) Non-interpolated motion signatures for two example long videos with all corresponding clips. Coloured outer boxes correspond to outlined boxes in A) right. C) First frames of each of the 125 video clips arranged according to the layout of A) right. Inset blue numbers correspond to the curves in B).

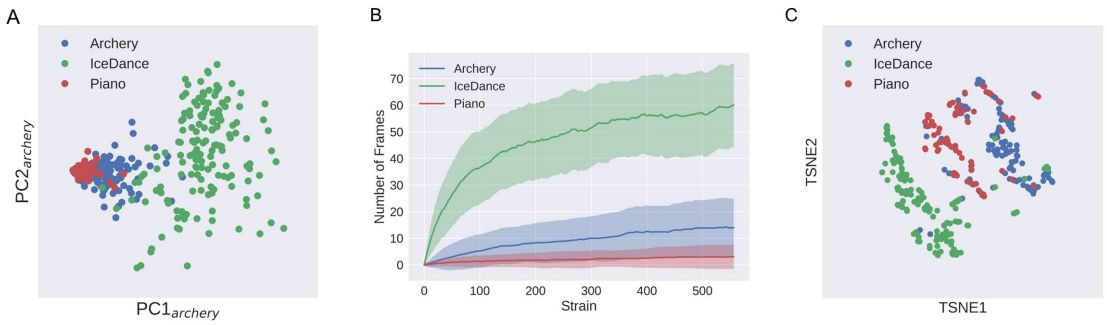

Figure 4.9: Comparing the motion of ice dancing and piano playing with archery. A) The ice dancing (n=158) and piano playing (n=104) video clips plotted onto the same axes defined by the two principal components of the archery videos. Each point is a videoclip. B) The average mesh strain curve for each action class (solid coloured lines). The respective coloured regions marks ±1 standard deviation of the mean curve. C) t-SNE embedding using the MOSES mesh strain curve of all video clips across the three action classes.

Experiment 3: Can MOSES be used as a Unique Signature?

The MOSES signature as defined in Eqn.(4.12) is meaningful in capturing the global motion patterns and can be used to assess 'outlier' videos of the same motion but how distinctive is this signature? To what extent can it distinguish one motion type e.g. archery to a different motion such as ice dancing or piano playing despite the inherent inter-variation within the same action class?

For biological studies it is common to compare the effects of perturbation (genetic or external factors) with respect to the reference wild-type condition. We construct a toy experiment that mimick this idea of identifying different phenotypes but with only initial knowledge of the wild-type variation. Using the archery videos as a 'wild-type control' we ask whether after setting the principal components (PC1, PC2) based on archery can we subsequently project ice dancing and piano playing MOSES signatures using the archery principal components to deduce these as two separate distinctive actions. In computer vision this task is also known as zero-shot learning, solving a task without having received any training examples of the task beforehand. Fig.4.9A shows that the inter-difference across actions using the MOSES signature is greater than the intra-difference within one action.

Further by the size of the 2D space mapped out by the points, the variation in motion was greatest in ice dancing compared to piano playing with archery in the middle. This separation appears primarily driven by the size of the motion alone, (Fig.4.9B). The application of t-SNE confirm the ability to largely separate the three distinct actions based on MOSES signatures though not fully for archery and piano playing. Most likely this is due to the loss of the spatial information in the MOSES signatures. Though somewhat artificial and qualitative, this experiment suggests that MOSES can be used as a potential signature for zero shot applications. Initial investigations building a CNN classifier using the forward MOSES motion signature on the full UCF101 dataset yielded an average classification accuracy of ∼20% across classes which is significantly better than random guessing (∼1%) confirming our observations of its informativeness but cannot compare to state-of-the-art results (c.f. 85%-95%, Table.2.4) which all heavily optimise spatiotemporal feature extraction using end-to-end learning approaches. This is not surprising as the proposed MOSES signatures average out the spatial motion variations across superpixels. Thus the reported classification accuracy is primarily due to temporal motion variations only. One would expect significantly improved results if we adopt similar strategies as that used for improved dense trajectories, (Wang and Schmid 2013); using alternative pooling strategies that do not average out the spatial motion component such as bag of words, including appearance-based features such as HoG or CNN features along the superpixel trajectories and removing camera motion prior to track extraction. For future work we plan to study architectures for incorporating the full spatiotemporal feature matrix of the dynamic mesh (c.f. Ch.4.3.2) with and without the connectivity along with appearance features (see chapter 7 future work for details).

4.6.2 Single cell tracking

The ability to track single cells is crucial to understanding individual cell motion and therefore single cell tracking algorithms has emerged as the *de-facto* tracking method.

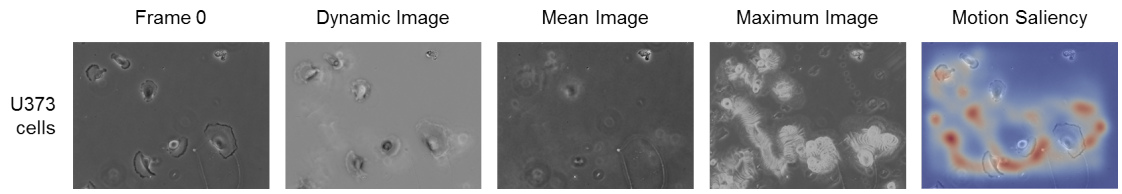

Figure 4.10: MOSES motion saliency applied to an example video of U373 cells from the cell tracking challenge dataset.

Much research efforts has been devoted to improving single cell tracking precision leading to the development of many algorithms such as Padfield et al. 2011; Schiegg et al. 2015 (reviewed in Ch.2.2). Despite their excellent reported performance however, these algorithms all require significant efforts in parameter tuning to obtain optimal performance on datasets they were not developed upon, from cell type to cell type, for image datasets from different acquisition setups or they are difficult to reimplement effectively. Theoretically, global motion patterns captures the movement of all single cells. Here we asked whether in certain scenarios if the global motion from dense optical flow combined with superpixels could be used as a simpler alternative to track single cells? Previously, superpixels (Wang et al. 2011b; Yang et al. 2014) have been investigated as a means to extract mid-level features for feature-based trackers for general object tracking and in Schiegg et al. 2015 as input to a graph cut algorithm for single cell tracking. But superpixel tracking with dense optical flow has not really been explored for cell tracking. In the process we were also able to assess the temporal continuity and specificity of the derived superpixel tracks. Unlike computer vision there are few annotated datasets for objective comparison of single-cell tracking and none that is large. Here we assessed MOSES on the training videos of the cell tracking challenge dataset of Maška et al. 2014 which has annotated 'ground-truth' single cell tracks across 5 different cell types with a mixture of fluorescent and DIC imaging, (Table.4.2). Instead of providing a single numerical metric of performance which begs the question of the most appropriate metric, c.f., Jaqaman et al. 2008; Luo et al. 2014; Maška et al. 2014; Leal-Taix et al. 2017, I chose instead to visually compare the derived single cell tracks to the provided ground-truth tracks for validation for each of the 10

Cells	Number of Videos	Objective	Time Step
MSC rat mesenchymal cells	2	10x/0.3	30 min
GOWT1 mouse stem cells	2	63x/1.4 oil	5 min
HeLa cells	2	10x/0.4	30 min
U373 glioblastoma-astrocytoma cells	2	20x/0.5	15 min
Simulated (Sim.) HL60 cells	2	40x/1.3 oil	29

Table 4.2: Table summarising the datasets used for MOSES single cell tracking from the cell tracking challenge, http://www.codesolorzano.com/Challenges/CTC/Datasets.html

unique videos in the entire dataset.

Fig.4.10 shows the first of two training videos for U373 cells, a glioblastoma-astrocytoma cell line which experience no cell division during the imaging duration but does exhibit significant shape changes and cell migration as evidenced by the maximum-projected image (but not reflected in the dynamic image). Whereas dynamic images may work well for normal RGB videos of everyday scenes, potentially the homogeneity of colour in DIC videos and the constant movement of the cells does not lend itself well to such a visualisation method. In contrast the MOSES saliency map is not affected and the maximum image also serves as a useful guide for visualising the video motion.

To produce single cell tracks, superpixel tracks were extracted as described (Fig.4.1) with 5000 superpixels. Individual cell segmentation of the initial frame was then used to assign superpixel tracks to each individual cell. For each cell, the longest track of those within a distance of $3w_s$ (w_s: average superpixel width) of the initial centroid is selected to represent the individual cell motion. The process is illustrated in Fig.4.11A for U373 cells with the corresponding result compared to

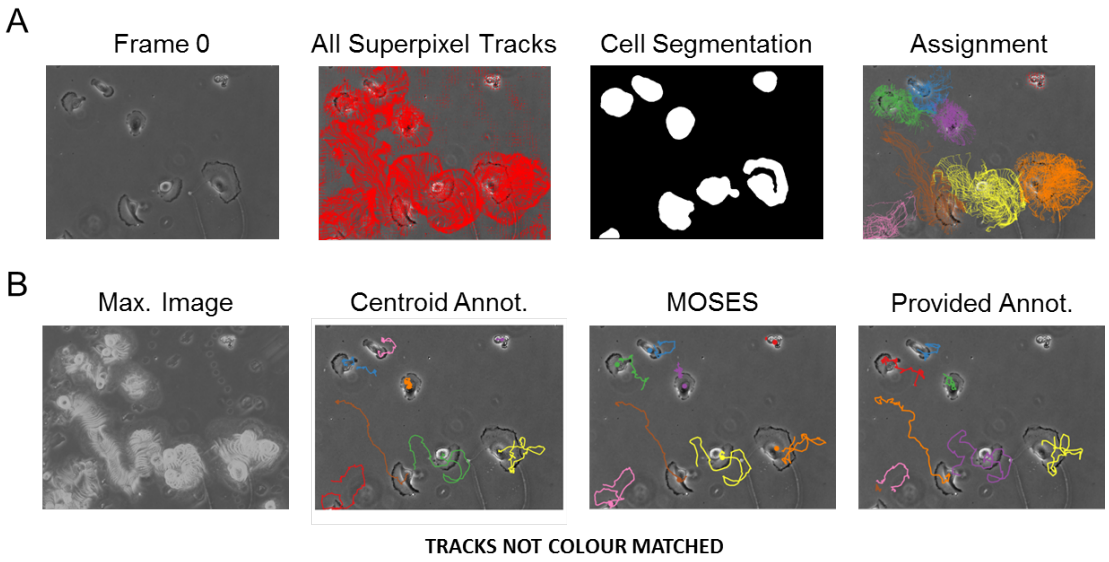

Figure 4.11: MOSES superpixel tracking of single cells. A) Overview of applying MOSES showing assignment of superpixels through coarse cell segmentation. B) Comparison of MOSES superpixel tracking with maximum projected image and manually annotations that were provided by the challenge and reannotated by a human annotator.

manual annotations, (Fig.4.11B). Since MOSES does not handle certain aspects of single cell tracking such as cell division, cell entry and exit from the field of view we do not expect pixel perfect performance. As such, for all displayed results, we compared MOSES tracks to annotated tracks that originate only from cells present in the initial field of view. In our experience the provided annotations (Provided Annot.) was not very accurate as can be seen when compared to the motion pattern revealed by the corresponding maximum image. For U373 cells, fellow PhD student Carlos reannotated the videos from scratch, identifying in each frame the cell nuclei. It can be seen that MOSES compares extremely well to the new centroid annotations even in the extremely challenging lower left border case in Fig.4.11B. The results for the remaining videos are shown in Fig.4.12, where if not indicated otherwise the manual annotations used for comparison were those that were provided. For U373, GOWT and MSC cells where their migration motion dominates over proliferation and we have correctly identified the location of initial cells through cell segmentation, MOSES performs visually well despite the very large shape deformations that U373 and MSC cells undergo, despite not having accounted for these changes explicitly in

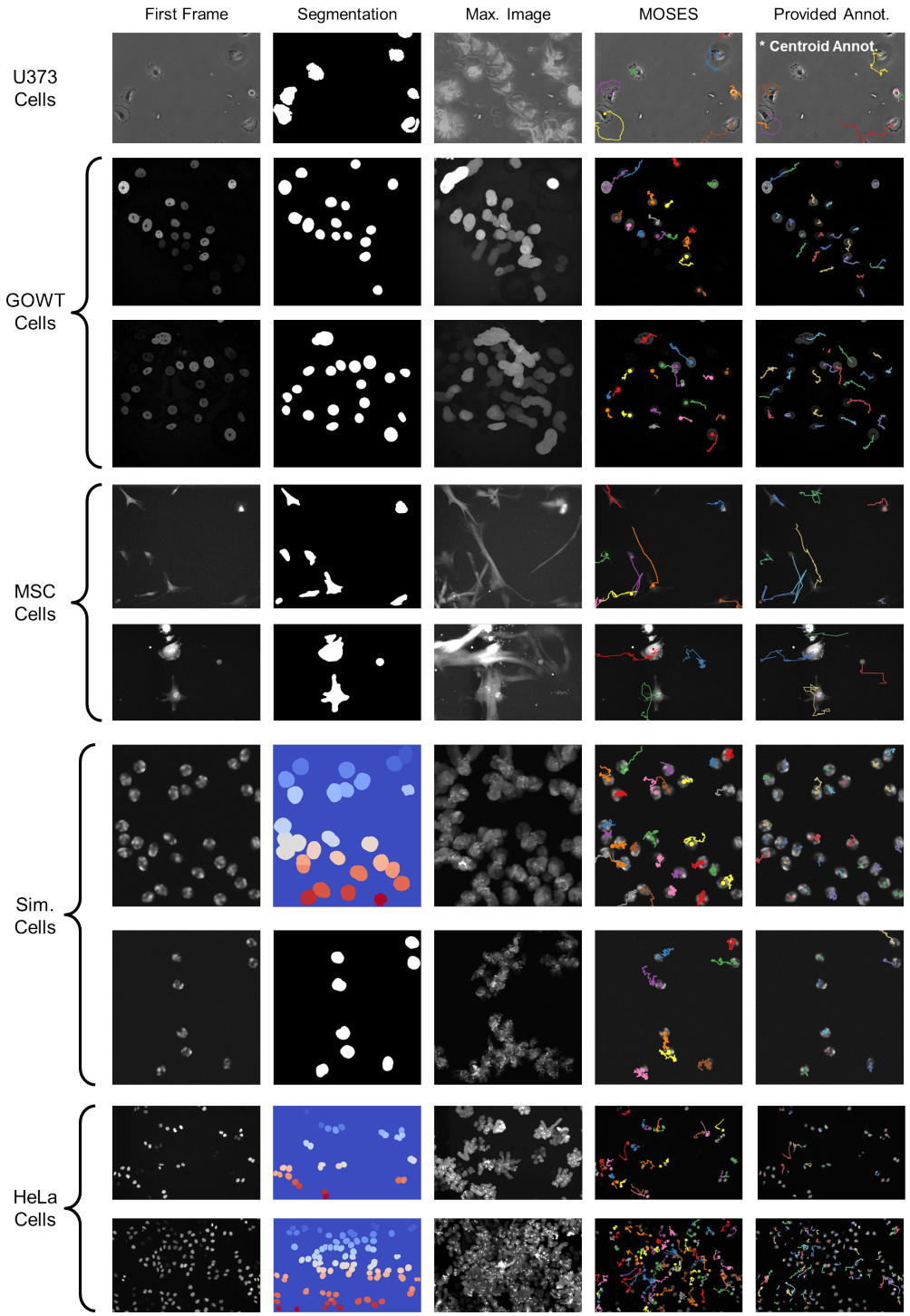

Figure 4.12: MOSES superpixel tracking applied to single cells. First frame is shown for all videos. Cells were segmented based on the entropy image computed from the pixel intensities. For segmentation masks that are coloured, watershed was applied to split cells.

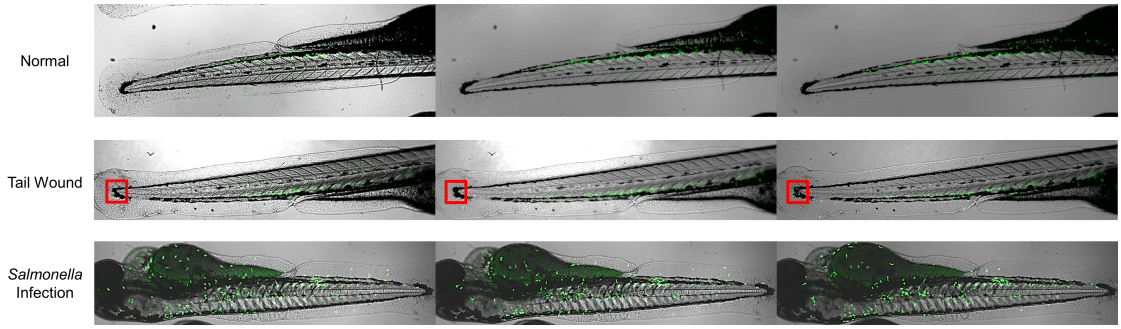

Figure 4.13: Different neutrophil motion behaviour as a result of wounding and bacterial infection. A tail wound was inflicted using the confocal microscopy laser 15mins from the start of acquisition for a total acquisition time of 120 mins. Red boxes highlight the tail wound. Wounded tails show enlargement and trigger local migration of neutrophils.

the motion extraction. Further the tracks produced are smoother and more realistic when compared to the provided annotations. In simulated (Sim.) and HeLa cells where the individual cells are not primarily migrating but mainly proliferating, with no explicit mechanism for distinguishing cell proliferation from cell migration, MOSES is prone to overestimate the biological motion. Whilst it is correct in assigning the global movement if we compare to the corresponding maximum image, due to the spatial proximity, the assigned motion most likely also includes the additional motion of the daughter cells and their cell divisions. MOSES thus could provide a quick solution to single cell tracking when cell proliferation is rare. Should proliferation be more dominant it would be more reliable to utilise a specialised single cell or particle tracker for cell tracking. In conclusion MOSES can robustly recover global motion patterns from different image modalities as demonstrated by the presented experiments above with spiked-in cells, UCF101 and single cell tracking. More objectively we could assess the robustness of MOSES to recover global motion patterns using generated ground-truth with simulation, by attempting to recover simulated global flow and associated streamlines at different noise levels. We now discuss two potential applications where MOSES could provide alternative insights compared to the normal single-cell tracking solution.

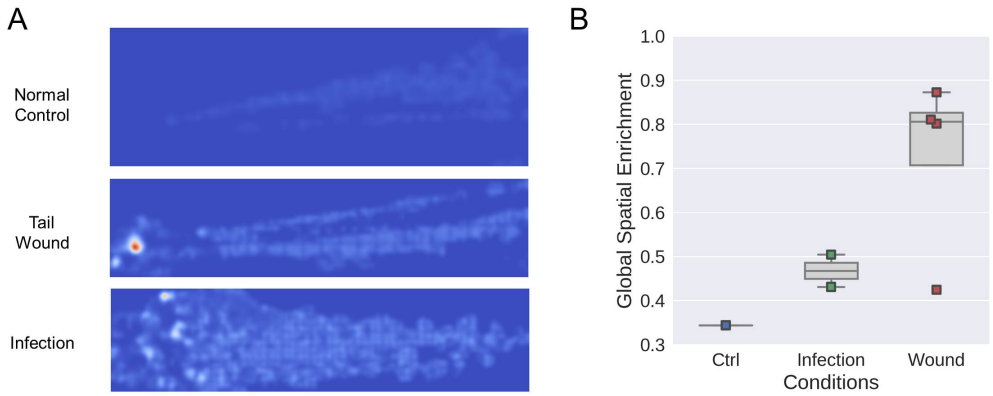

Figure 4.14: Different neutrophil motion behaviour as a result of wounding and bacterial infection. A) Motion saliency heatmaps for a typical video from each condition with 10000 superpixels and the radial density mesh. B) Global spatial enrichment of each condition based on the motion saliency. This value ranges from 0-1.

4.6.3 Monitoring Immunosurveillance in Zebrafish

Single cell tracking is an extremely difficult problem that potentially can only be solved exactly for certain situations given optimal imaging conditions. Often however for biological applications it is desired only to show differences between two experimental conditions to aid experimental protocol development and hypothesis generation. Seeking exact solutions in these cases to single cell tracking could quickly become more challenging and lead to far more complex downstream analysis than might actually be required for answering the scientific question of interest. One such example that benefits more from a global perspective of movement is immunosurveillance in zebrafish using 2D fluorescence microscopy where we desire to be able to distinguish the different neutrophil motion behaviour in response to external stimulation such as bacterial infection or wounding. Quantitative tracking of neutrophils *in-vivo* in zebrafish is not standard practice due to the low-resolution acquisition. Currently researchers manually count the number of neutrophils and manually track the neutrophils (Li et al. 2012). If automated tracking is applied, it is usually in a small region of few neutrophils (Henry et al. 2013; Robertson et al. 2014). In an ongoing collaboration with the Gyrd-Hansen Lab we have two groups of zebrafish that have been externally stimulated, (Fig.4.13). For the first group a specific tail wound was inflicted 15 mins after the start of filming and was filmed

for a total of 120 mins at 10 sec intervals. For the second group the zebrafish was chronically infected by pre-injection prior to filming with *Salmonella* bacteria into the circulatory system. In both cases only the neutrophils were fluorescently labelled with GFP. The challenge is to distinguish the different patterns of behaviour for the different conditions.

Using MOSES we explored the idea of phenotyping the resulting motion patterns of the circulating neutrophils across the whole zebrafish (GFP-tagged and green) using the motion saliency map with 10000 superpixels and the radial density mesh. The most sensitive mesh, the radial density mesh was chosen from the described four meshes (Ch.4.3.2) to maximise the signal from the small number of local neutrophils that exhibit directional movement following wounding. For the normal control, spatially the motion saliency heatmap with forward tracks reveal no spatial motion enrichment, for the tail wound, it reveals very precisely significant local enrichment of motion at the site of wounding and for infection it reveals globally elevated motion enrichment, (Fig.4.14A). A simple way to summarise the spatial enrichment using a single statistic is to compute a normalised signal-to-noise index after partitioning only the motion saliency map within the area of the zebrafish into 'high' and 'low' pixel intensity values with Otsu thresholding:

$$\text{Global Spatial Enrichment} = \frac{\mathbb{E}[\text{High}] - \mathbb{E}[\text{Low}]}{\mathbb{E}[\text{High}]} \tag{4.19}$$

where $\mathbb{E}$ is the mean. The proposed normalised signal-to-noise ratio, (Fig.4.14B) corroborates with the qualitative observational findings of (Fig.4.14A). One may note an outlier in the wounded condition. This is because for that video, 3 local heatspots similar to the wound heatspot in Fig.4.14A was present and the proposed simple index does not capture this spatial multimodality. Extensions of this index can be proposed that directly measure the spatial distribution such as peak finding approaches.

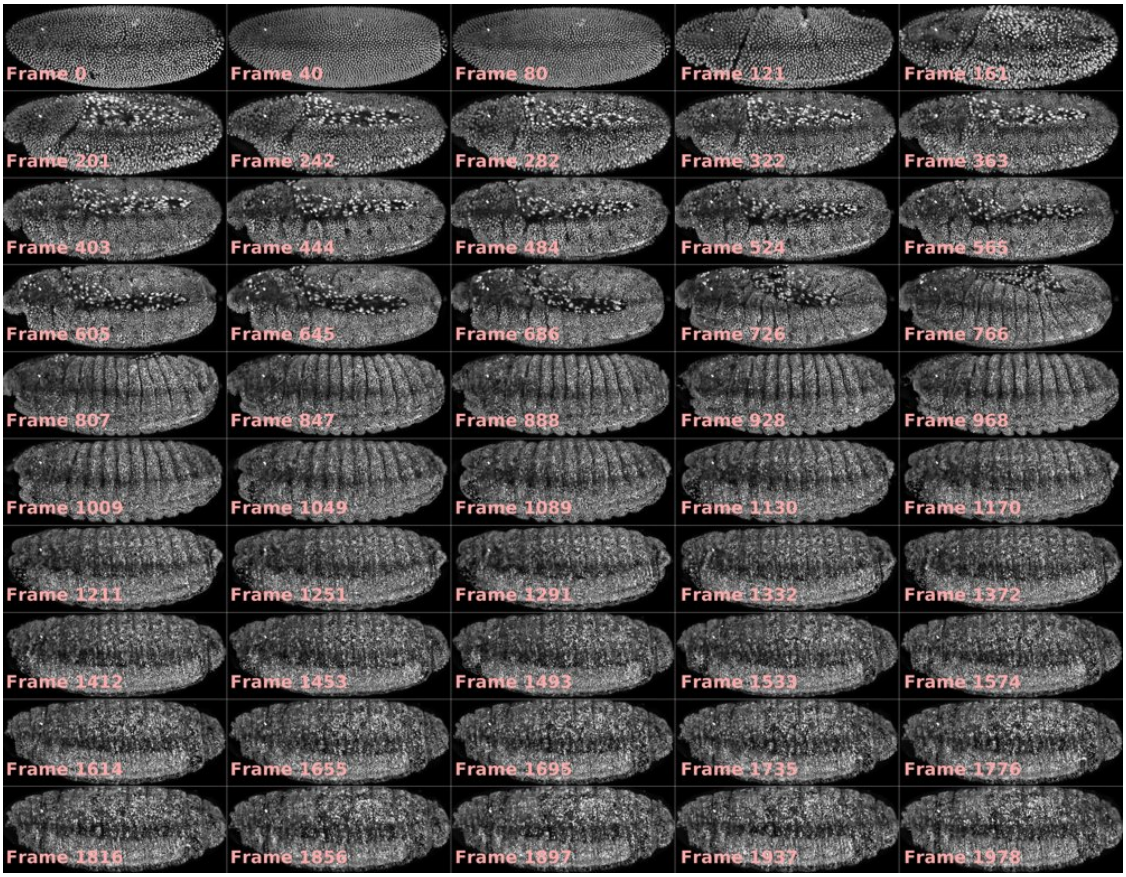

Figure 4.15: Video snapshots of *drosophila* embryo development. Static frames were regularly sampled from supplementary movie 3 of Tomer et al. 2012 ventrolateral view. It comprises a total of 1979 frames a period of 19.5 hours, recorded at 35s intervals.

4.6.4 Monitoring developmental processes

One of the major outstanding problems for *in-vivo* imaging is the analysis of longtime acquisitions particularly where cells may undergo extensive motion dynamics with cell populations disappearing altogether from the field of view. Using the example of a developing *drosophila* embryo, we illustrate how MOSES and global tracking can be used in an effective manner to quickly make insights as a complementary analysis suite to single cell tracking for understanding the underlying complex motion dynamics.

MOSES enables long-time tracking of cell populations

We analysed Supplementary movie 3 from Tomer et al. 2012 with MOSES. This is a 2D maximum intensity projected sequence of the full 3D acquisition (Fig.4.15).

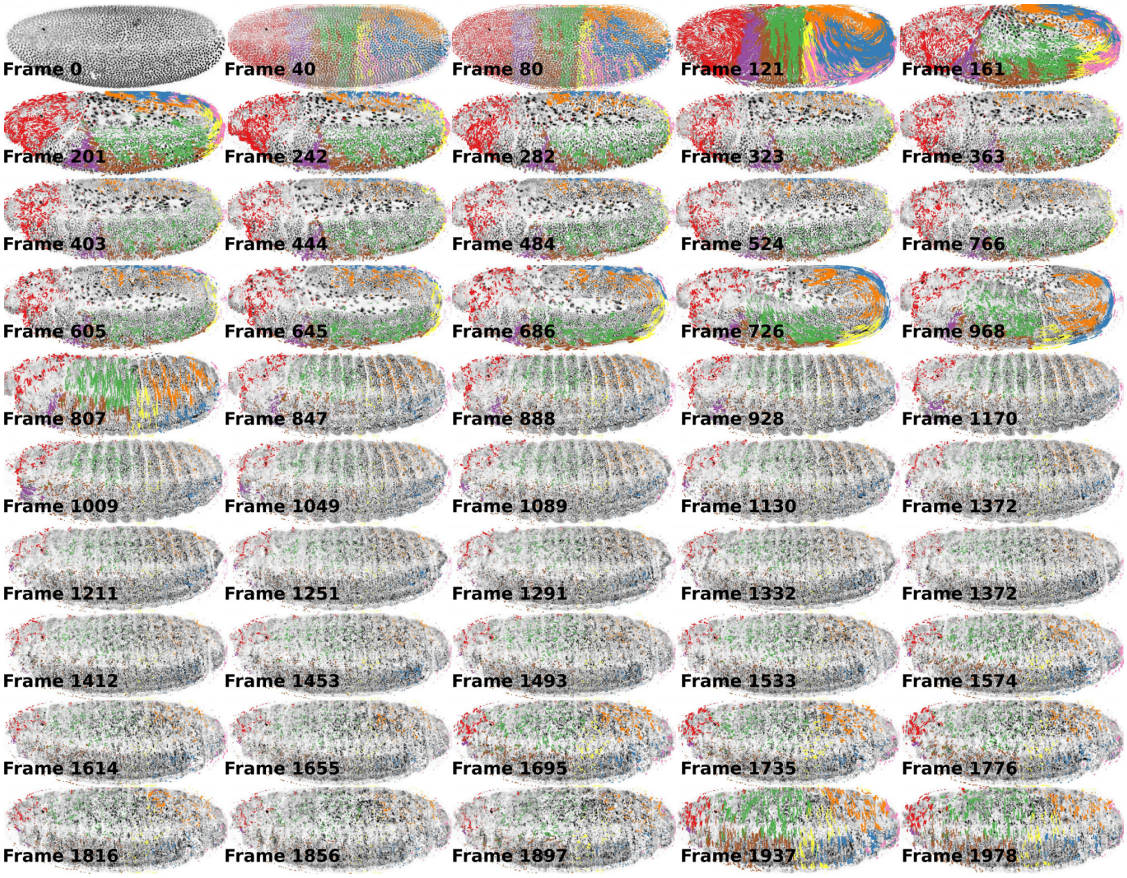

Figure 4.16: Superpixel tracks of *drosophila* embryo development. Corresponding superpixel tracks with 10000 superpixels of Fig.4.15. For frame $t > 0$ tracks are plotted from $t - 10$ to time t. Tracks were clustered with a GMM model after PCA dimensional reduction with 8 components.

As such, the perceived motion will be biased towards the visible cells closest to the camera. Movement of cells in the bulk will not be captured. Nevertheless due to the axial symmetry of the motions present, the 2D projected video is sufficient for revealing global motion patterns.. Superpixel tracks were extracted with 10000 superpixels. To aid visualisation of the resultant very dense superpixel tracks, track clustering (Ch.4.4.1) was applied using the entire trajectory history over all frames with 8 clusters, (Fig.4.16). The number of clusters was arbitrarily selected to visualise the different types of motion patterns present. In general an automatic model selection method such as BIC (Bayesian Information Criterion, (Fraley and Raftery 1998)) or silhouette analysis (Rousseeuw 1987) can be used to select an appropriate number of clusters. As seen by the temporal snapshots

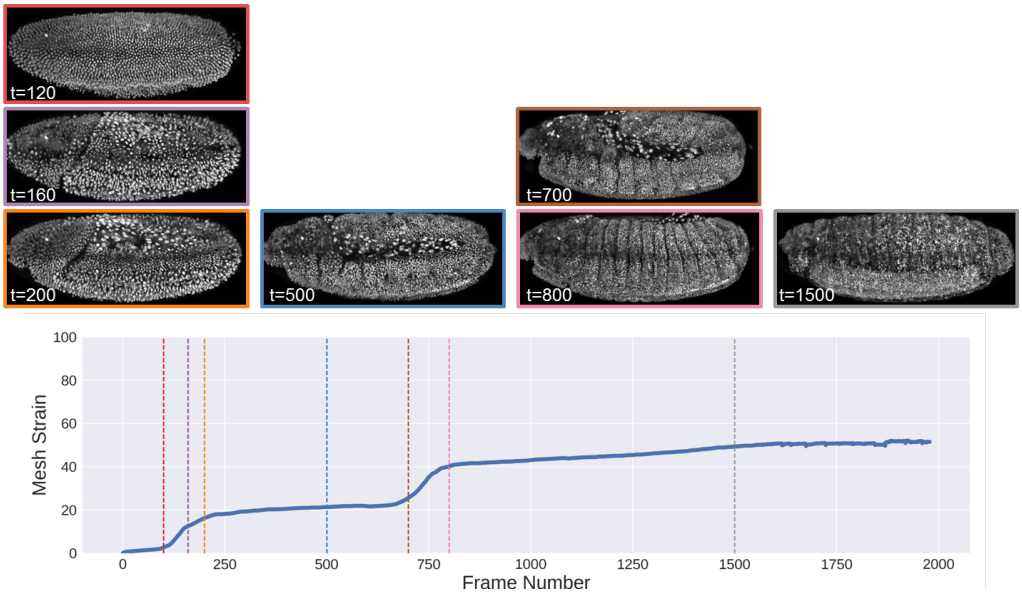

Figure 4.17: MOSES mesh strain analysis of *drosophila* embryo development.

where the tracks at frame t are plotted from time $t-10$ to t, we see how tracks of the same cluster track local motion patterns that appear to share the same 'fate', highlighting the same consistent group of superpixels (cells) as they move and end in a similar spatial region.

MOSES mesh strain automatically identifies distinct motion regimes

The MOSES superpixel tracks can consistently follow the cellular motion over long times as demonstrated through clustering but as clearly seen from the video and track snapshots (Fig.4.15, 4.16) there appears to be distinct temporal motion patterns, for example an upwards swirling motion between ∼frames 121-241 (Fig.4.15) and ∼frames 160-200, (Fig.4.17). Can MOSES be used to automatically identify these regimes for more detailed study such as single cell tracking?. We tested this by computing the forward MOSES motion signature or mesh strain curve (Ch.4.4.2). Fig.4.17 shows how the forward MOSES motion signature automatically identifies large motion regime transitions in particular at frames 120-250 the large upwards rotational development of the front and back and at frames 700-800 the rotational movement and reorientation of the back through increases in the mesh strain. Manual human annotation identifies 4 motion regimes. In addition to 1) the large

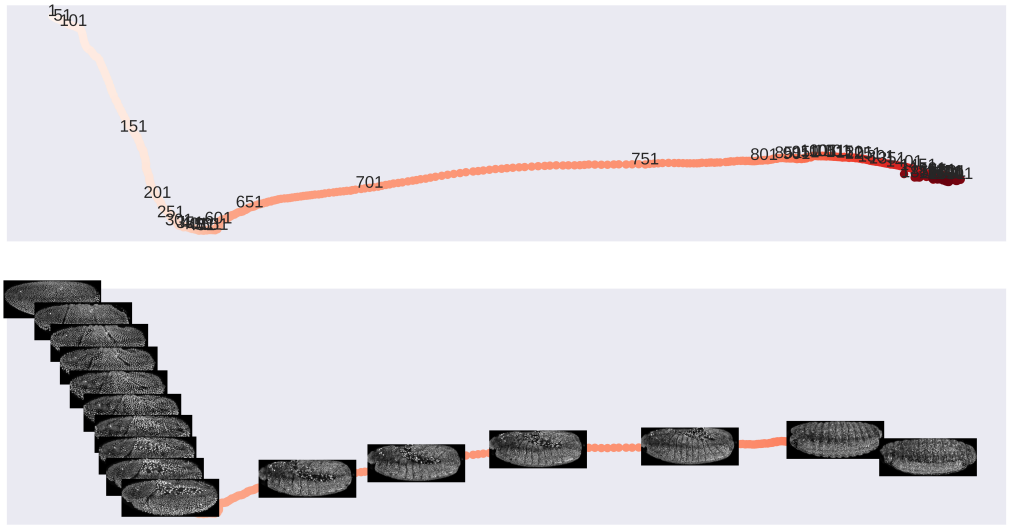

Figure 4.18: MOSES frame embedding of *drosophila* embryo development. 2-component PCA with whitening was used to project the video frames into 2D using the MOSES frame signature, Eqn.(4.15). Top: PCA embedding of the MOSES frame signature with points coloured chronologically by increasing redness. Frames are labelled in increments of 50 frames. Bottom: automatically sampled video frames selected to ensure a minimal distance between successive frames for viewing.

upwards rotational development of the front and back at frames 120-250 and 2) the rotational movement and reorientation of the back at frames 700-800, there are two subtler temporal patterns involving less movement, 3) sideways embryo movement at frames 1500-1750 and 4) a jitter at frames 1900-2000 (which could also be due to the mounting). The latter two is much less obvious in the MOSES strain curve, pattern 4 is still just about detectable. An alternative interrogation method is through frame embedding of the frame signature, (Eqn.(4.15)). Fig.4.18 illustrates how the application of PCA automatically clusters together visually similarly looking video frames whilst spreading out frames with very dissimilar appearance. The overall effect is amplification of the large motion transitions seen in the MOSES strain curve of Fig.4.16, allowing easy identification of the 2 dominant motion regimes. One immediate application of this would be the construction of key frames to summarise the video motion. Video frames in areas of high point density in the frame embedding can be summarised using one frame whilst areas of low point density require multiple frames.

4.7 Summary and Conclusions

- MOSES uses dense optical flow to derive long-time superpixel tracks. Tracks are joined to form dynamic meshes. Mesh-based statistics incorporating collective motion and individual track statistics together comprehensively characterise cellular motion (Ch. 5, 6). The meshes enable i) high-level motion signatures for video clustering and ii) motion saliency maps to highlight spatially distinct motion patterns.
- MOSES was demonstrated by application to different datasets:
    1. **UCF101 Action recognition** - Motion saliency maps accurately spatially localise motion sources despite background motion. Motion signatures quantify heterogeneity of videos from same action class with PCA. New actions are distinguished without seeing new examples (zero-shot learning). Frame motion signatures organise video frames temporally and semantically using PCA.
    2. **Single cell tracking challenge** - Superpixel tracks accurately recover single cell tracks in the presence of no proliferation.
    3. **Zebrafish immunosurveillance** - Motion saliency maps enable novel characterisation of neutrophil motion patterns even for low-resolution videos without need for explicit cell tracking.
    4. ***Drosphila* Embryo Development** - Consistent long-time tracking of motion patterns. Motion signatures automatically temporally localise the distinct motion regimes present in the video.
- MOSES is extensible to 3D and nD using multidimensional optical flow. The dynamic mesh connectivity requires only a definition of distance.
- **Limitations**
    1. Fixed number of superpixels. Tracking density for large motion leads to insufficient coverage e.g. *Drosophila*. Solution: online introduction and tracking of new superpixels.
    2. No distinction of cellular processes e.g. cell proliferation, cells entering and exiting. Single cell tracks are overestimated in proliferation dominant videos. Solution: use dedicated single cell tracking.
    3. Current superpixel averaged motion signatures do not encode sufficient spatial motion variations. Videos with different lengths give different length signatures and is difficult for classifiers. For UCF101, we preliminarily achieve $\sim 20\%$. This is future work (see Ch.7).

5

Phenotyping Cell Population Interactions

Contents

In this chapter we show how the Motion Sensing Superpixels (MOSES) framework can be used to systematically capture diverse features of cellular dynamics from videos for motion characterisation of subtle epithelial population interactions. All experiments and timelapse imaging in this chapter were performed by Carlos Ruiz-Puig.

5.1 Barrett's Esophagus, Esophageal Adenocarcinoma and the Squamous-columnar Junction

An example application problem that benefits from a high-throughput computational analysis is the formation of stable boundaries between homo- and heterotypic cell populations. When two cell populations meet *in-vivo*, they often form a sharp, stable interface termed a 'boundary', with limited intermingling (Dahmann et al. 2011). In adult humans, sharp boundaries separate different types of epithelia: for example between the squamous and columnar epithelia in the esophagus, cervix and anus. Disruption of these boundaries can lead to disease. Disruption of the squamo-columnar epithelial boundary in Barrett's Esophagus (BE) confers a 30-50 fold increased risk of esophageal adenocarcinoma (EAC) (Gaddam et al. 2013). Understanding how tissue dynamics relates to pathological phenotypes and how it can be affected by intrinsic and extrinsic factors is therefore a key issue.

5.2 In-vitro model to study the spatiotemporal dynamics between different cell populations

5.2.1 Temporary Divider Co-culture Assay

To develop a motion analysis framework an experimental model was established to assess boundary formation *in-vitro* between squamous and columnar epithelia found at the esophageal SCJ (squamous-columnar junction) (Fig.5.1A). Three epithelial cell lines were used: EPC2 (an immortalised squamous epithelial cell line from the normal esophagus (Harada et al. 2003); CP-A (an immortalised BE cell line with columnar epithelial properties (Merlo et al. 2011); and OE33 (derived from EAC

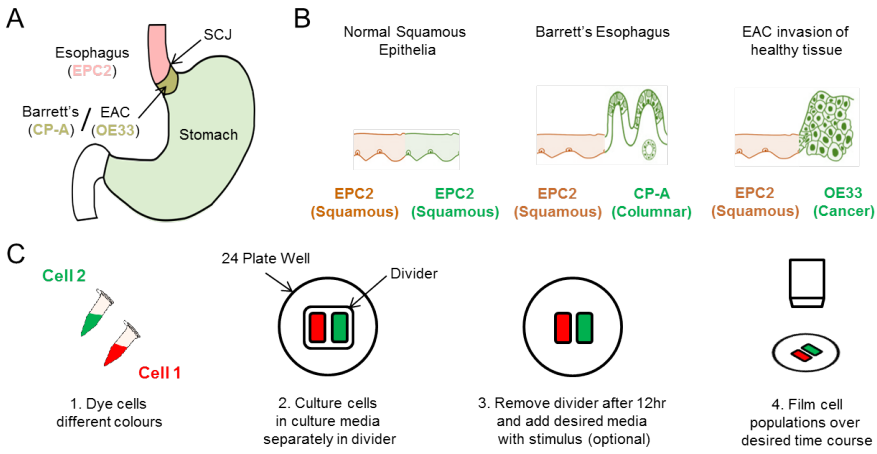

Figure 5.1: Temporary divider system to study interactions between cell populations. A) The squamous columnar junction (SCJ) divides the stratified squamous epithelia of the esophagus and the columnar epithelia of the stomach. Barrett's esophagus is characterised by squamous epithelia being replaced by columnar cells. B) The three main epithelial interfaces that occur in BE to EAC progression. Pictures were adapted from figure 2 of Evans et al. 2016. C) Overview of the experimental procedure, described in steps 1-3. In our set-up cells were allowed to migrate and were filmed for 4-6 days after removal of the divider (step 4).

(Boonstra et al. 2010). To model the interfaces that occur in the esophagus we used the combinations: EPC2:EPC2 (squamous:squamous, as a normal control), EPC2:CP-A (squamous:columnar, as in Barrett's esophagus) and EPC2:OE33 (squamous:cancer, as in EAC) (Fig.5.1B). In this experimental model (Fig.5.1C), two epithelial cell populations are co-cultured in the same well of a 24-well plate in media with 5% serum (FBS), separated by a divider with width 500μm (Ibidi). The divider is then removed after 12hr and cells allowed to migrate towards each other. Each cell population is labelled with a lipophilic membrane dye (Celltracker Green (CMFDA, Life Technologies) and Celltracker Orange CMRA, Life Technologies), which provides uniform staining, low phototoxicity and can be used *in-vivo* and to label primary cells (Progatzky et al. 2013). We tested the effects of the dye on proliferation and migration in two populations of EPC2, filmed over 96 and 144 hrs.

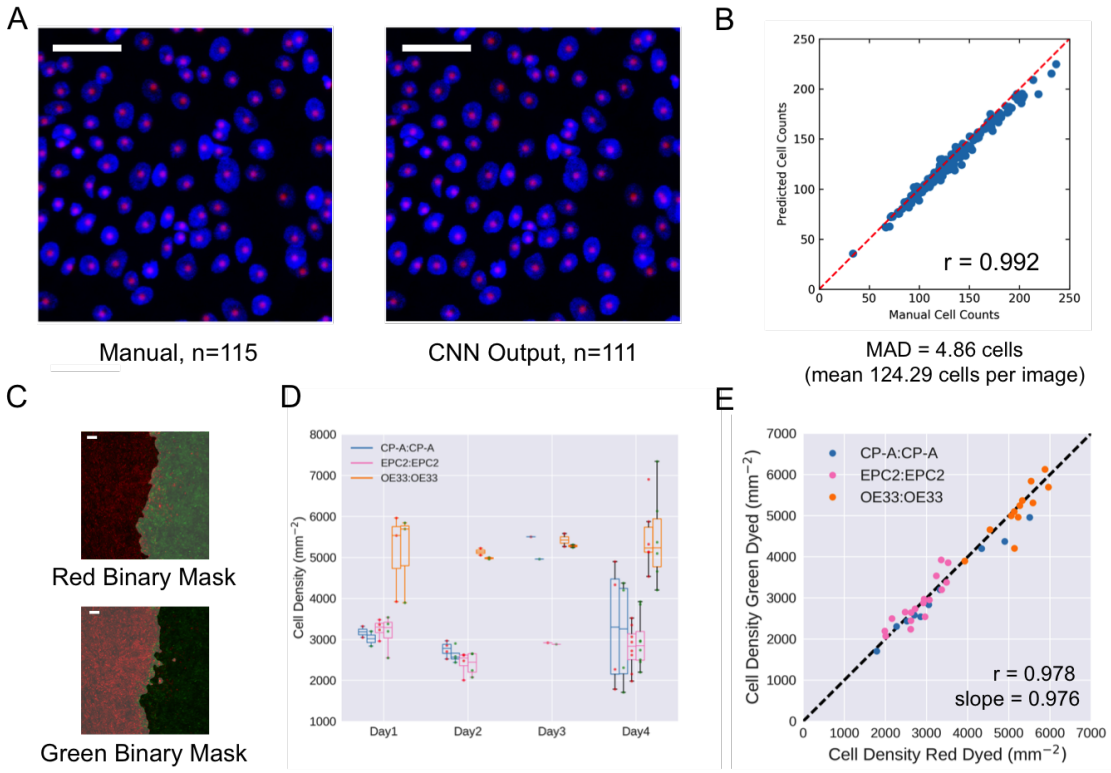

Figure 5.2: Assessment of cell proliferation with dye. A) Automated cell counting with convolutional neural networks (CNN). An example of a 64×64 pixels image patch of cells stained with DAPI (blue), with individual cells counted manually (left) or by automatic counting (right). Red spots mark individual counted cells. B) Plot of manually annotated cell counts vs automated cell counts tested on 64×64 image patches (n=200). Each point is a patch. The mean absolute deviation (MAD) was 4.86 cells, a percentage error of 3.91% for an average of 124.29 cells per image. Pearson correlation coefficient, r=0.992. Red dashed line is the ideal identity line. C) Image segmentation of epithelial sheets coloured with red and green dyes, used for sheet specific cell counting. Grey shaded area is the excluded image area. D) Paired boxplots of cell density (number of cells/sheet area) for each monolayer from fixed samples collected at different times (up to 4 days (96hr) after divider removal). In each pair, the left boxplot is for the red labelled cells and right for the green labelled cells, indicated by red or green dots respectively. Each dot represents the value from a confocal image. Outline box colour indicates the cell type (see legend). E) Cell density of red vs green dyed cells in the same culture pooled from D) co-plotted on the same axis. Each point is a separate image. If a point lies on the identity line (black dashed), within the image, red and green dyed cells have the same cell density. All scale bars: 200μm.

5.2.2 Assessment of Proliferation and Migration with Dye

Proliferation Assessment

To assess proliferation, confocal images were taken of fixed samples at 0,1,2,3,4 days and stained with DAPI to highlight nuclei. A fully convolutional neural network (CNN) was trained in the manner of Xie et al. 2016 to automatically count the number of cells. Given a DAPI image the trained CNN outputs a cell density heatmap whose sum of pixel values equals the number of cells in the image, (Fig.5.2A). To train the network, cells were first manually annotated in 192 image patches (256×256 pixels) extracted from the original images (4096×4096 pixels), (Fig.5.2A). For each of the 256×256 pixel image patches, 50 randomly sampled 64×64 pixel patches were extracted to yield a training set of 10,000 images. A 70:30 train-test split was used to train the CNN for 200 epochs, batch size 100 with RMSprop (lr=0.001, rho=0.9, epsilon=1e-08, decay=0.0) in Keras (Theano backend) with MAE (mean absolute error) loss. The final test accuracy is reported on the result of applying the trained CNN to the original 256×256 labelled image patches, (Fig.5.2B) yielding a strong Pearson correlation of 0.992 with a mean absolute deviation (MAD) error of 3.9% for an average of 124.29 cells per 256×256 image. To count red/green dyed cells within a single image, epithelial sheets were segmented using their individual colour channel intensity images with K-means clustering (K=3) after downsampling the image by a factor of 4 (for speed). The resulting binary mask (retaining the two classes of highest image intensity) was cleaned by removing small objects (<200 pixels), filling holes and retaining the largest connected component before being upsampled back to its original resolution (4096×4096 pixels). To count, the respective binary mask is multiplied with the CNN density output image and summed over all the pixels, (Fig.5.2C).

From the DAPI images, the green and red fluorescent labelled EPC2 cells proliferated similarly with identical medians between the red and green dyed (red and green points in Fig.5.2D) within the interquartile range of the boxplots exhibiting the same cell density (number of cells divided by area) over the time

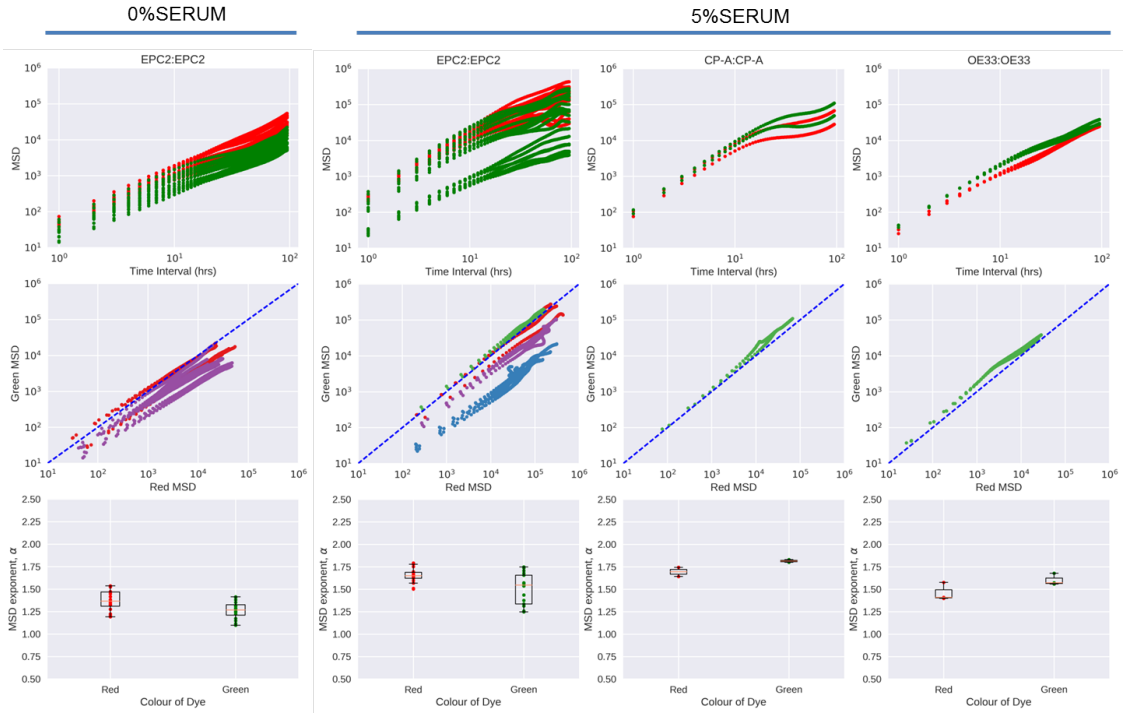

Figure 5.3: Cell migration is not affected by dye colour. (Top:) Mean squared displacement (MSD) curves computed from optical flow plotted on a $\log_{10}$-$\log_{10}$ axis as a function of the time interval. Lines are coloured red or green depending on the dye colour. Cell types are indicated above the plots. (Middle:) Each line shows the green vs red MSD curve from the same video. Lines are coloured by the experiment (different batches) in which videos were collected. The dashed blue line is the ideal identity curve. (Bottom:) Extracted exponent α from MSD$\propto \Delta t^{\alpha}$ for time intervals $0 < \Delta t \leq 20$ hrs, for the respective red and green sheet in one video. Each point is a video.

course. This is further confirmed by plotting red vs green cell density after pooling together all samples irrespective of day, (Fig.5.2E) with slope=0.976, Pearson correlation coefficient 0.978). Thus the dye has minimal impact on proliferation. The proliferation were similarly checked in CP-A and OE33, (Fig.5.2D,E) with the same conclusion.

Migration Assessment

To assess migration, the mean squared displacement (MSD) (Ch.3.1.1) as in Park et al. 2015 was computed:

$$\mathrm{MSD}(\Delta t) = \langle |\mathbf{r}_i(t + \Delta t) - \mathbf{r}_i(t)|^2 \rangle \tag{5.1}$$

where $\mathbf{r}_i$ is the (x, y) position of superpixel i at time t, Δt is a time interval and $\langle \cdot \rangle$ is the ensemble average over all time t and all superpixels i. For small time intervals Δt, the MSD increases as a power law:

$$\text{MSD}(\Delta t) \propto (\Delta t)^{\alpha}, \quad \text{if } \Delta t \text{ is small} \tag{5.2}$$

where α is a constant exponent. On a $\log_{10} - \log_{10}$ axis, this relationship is linear and of the form $y = mx + c$,

$$\log_{10} \text{MSD}(\Delta t) = \alpha \log_{10}(\Delta t) + \log_{10} C$$

The exponent α can therefore be determined empirically via linear regression. When α is unity, $(\alpha = 1)$ the movement is uncorrelated random Brownian motion and cellular motion is diffusive. When $\alpha > 1$, cellular motion is super-diffusive and when $\alpha = 2$, motions are 'ballistic'. Computing the MSD curves for EPC2:EPC2, the independent curves of the red and green cells have similar profiles (Fig.5.3 top panel) and this is confirmed by the strong positive correlation when the red MSD curve is plotted against the green MSD curve of the same video, (Fig.5.3 middle panel). A systematic deviation from the ideal blue dotted identity line is evident in Fig.5.3 middle panel due to an asymmetric positioning of the gap in that particular experiment, (c.f. Table.5.2). It is a batch effect. The MSD exponent $\alpha > 1$ inferred from $0 \leq \Delta t \leq 20$ hrs shows that both red and green cells exhibit the same type of super-diffusive migration, (Fig.5.3 bottom panel). The dye thus has minimal impact on migration. The migration were similarly checked in CP-A and OE33, (Fig.5.3) with the same conclusion.

5.2.3 Different Media, Collective Motion and Boundary Formation

We tested different combinations of the three epithelial cell lines, initially in serum-containing media (5% FBS) (Fig.5.4). In all combinations both populations moved together as a cohesive epithelial sheet. In the squamous EPC2:EPC2 combination, cells met and coalesced into a monolayer. In the squamous-columnar EPC2:CP-A

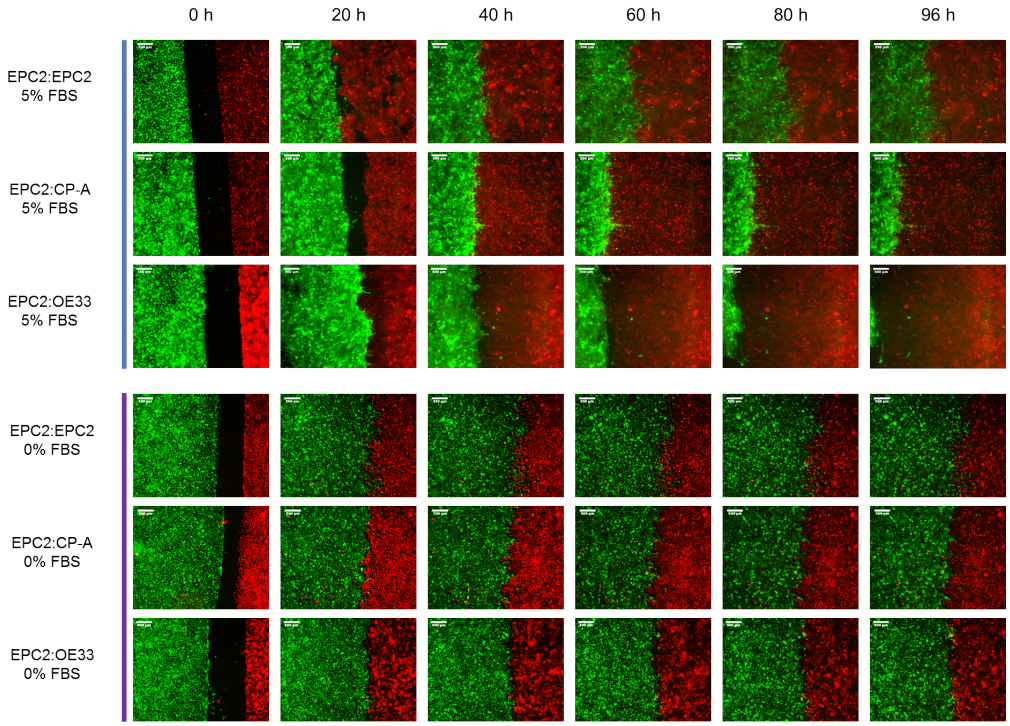

Figure 5.4: Video snapshots of cell combinations in 0% and 5% serum. The three combinations of epithelial cell types (EPC2:EPC2, EPC2:CP-A, EPC2:OE33) were filmed for 96hrs. Top: typical behaviour in serum and bottom: without serum. Scale bars: 500μm.

combination a stable boundary formed between the two populations after 72 hrs, following a short period of CP-A pushing EPC2, (Fig.5.4 5% FBS videos). In the squamous:cancer EPC2:OE33 combination, the cancer cell line OE33 pushes EPC2 out of the field of view. These phenotypes are consistent and reproducible, though in the latter combination of EPC2:OE33, it was interesting to find a rare number of videos (<10%) show the EPC2 cell line moving away from the OE33 cell line before the leading OE33 cell makes contact. It is not clear why the supposedly 'normal' EPC2 cells should behave in such a manner. If true, it could be an interesting cell-autonomous mechanism. For the purposes of the thesis and the development of MOSES, we did not extensively pursue the underlying biological mechanisms and include these videos as ultimately the EPC2 cells move out of the field-of-view, the same endpoint as the majority case of OE33 physically 'pushing' EPC2 out.

Evidence from systems including *Drosophila* embryonic parasegment (Monier

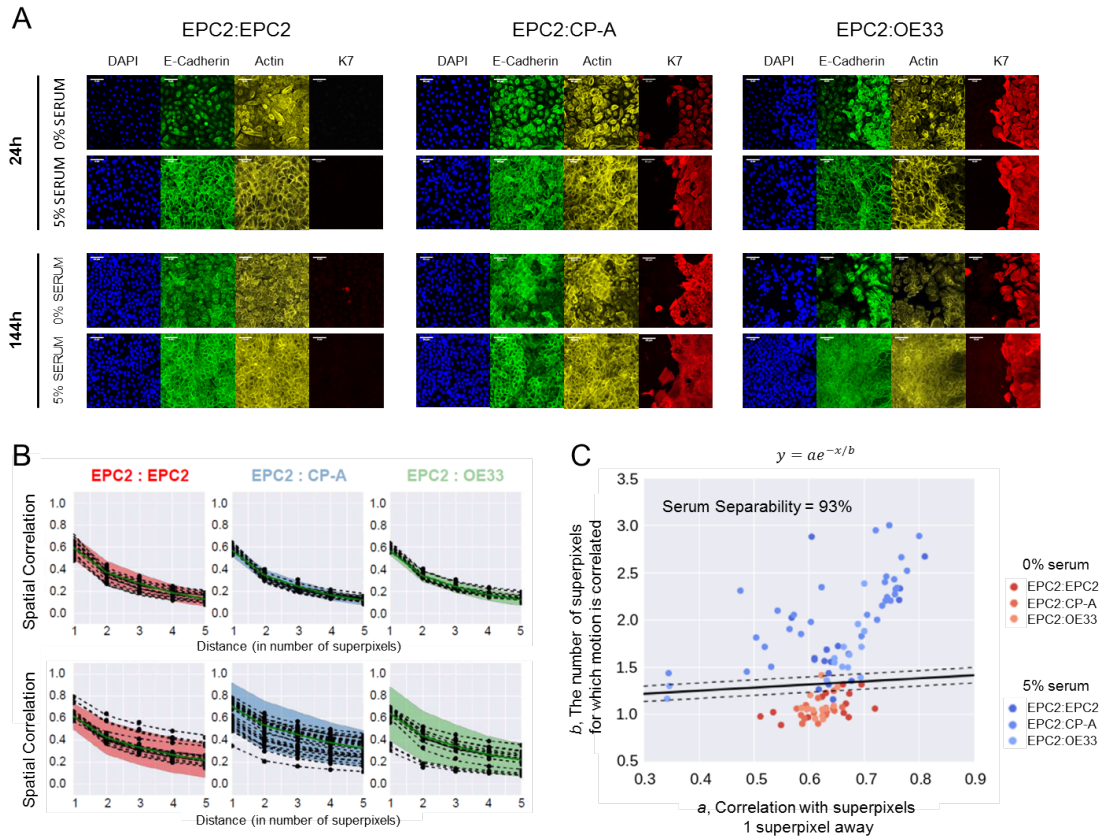

Figure 5.5: Collective sheet migration dynamics is lost in 0% serum. A) Samples were fixed for staining at 24h when the gap between the two sheets has just closed, and at 144h at the end of filming. K7 (red) is a specific marker of columnar cells. E-cadherin marks the adheren junctions. Actin (β-actin) marks the actin-cytoskeletion. All scale bars: 20μm. B) Spatial correlation curves computed from superpixel tracks as a function of superpixel distance. Black dots=computed values. Dashed black line=fitted line to black dots of the form $y = ae^{(-x/b)}$. Green solid line=median line computed from the black dots. Shaded region=$\pm$2 standard deviations of the green median line. C) Plot of the extracted values a vs b, where a is the correlation with superpixels 1 superpixel away and b the number of superpixels away for which motion is correlated. Each video is a point, see legend for colour code. The higher points are on the plot, the greater the collective motion. Black solid line is the support vector of a linear support vector machine (SVM) trained to separate 0% and 5% serum according to the values of a and b. Dashed black lines mark the SVM margin. Serum separability, the ability to predict if a video contains 0% or 5% serum is defined as the training SVM accuracy using the whole dataset, n=125 videos (n=16 each for EPC2:EPC2, EPC2:CP-A, EPC2:OE33 in 0% serum and n=17 (EPC2:EPC2), 30 (EPC2:CP-A) and 17 (EPC2:EPC2) in 5% serum.

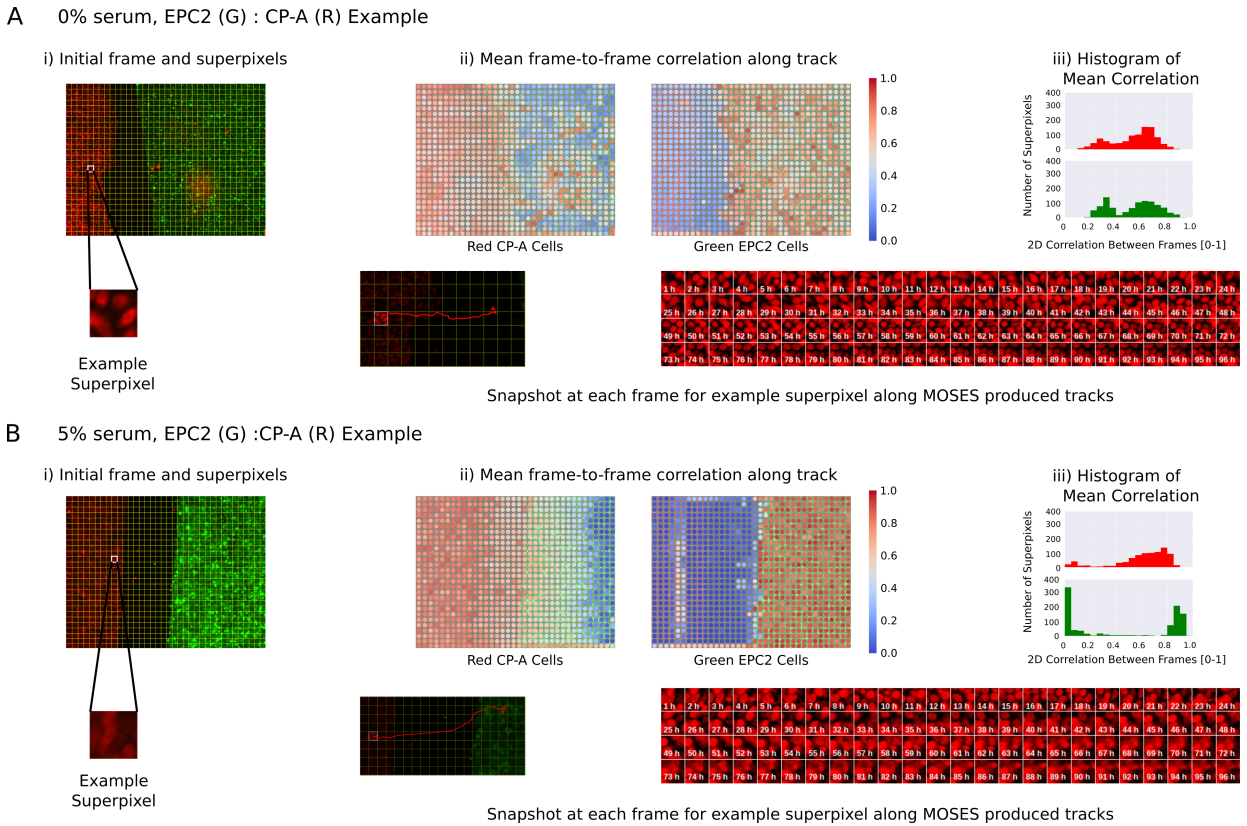

Figure 5.6: Reduced frame to frame correlation within tracked superpixels in 0% serum. A) Representative 0% serum EPC2(G):CP-A(R) example. i) Initial video frame with derived 1000 superpixels overlaid. For a single superpixel we extract the image patch it 'sees' along the computed superpixel track at each frame from 1-96h. The visual similarity of the image patch at the current time is compared to that of the previous time using the maximum value of the 2D normalised cross-correlation. For a single superpixel, this correlation is averaged to produce a mean frame-to-frame correlation from 0-1. ii) Spatial scatterplot of the initial positions of all superpixels coloured according to their mean frame-to-frame correlation for red CP-A and green EPC2 cells respectively. iii) Corresponding histogram distribution of the mean frame-to-frame correlation from A ii). B) Representative 5% serum EPC2(G):CP-A(R) example with i), ii) iii) showing the same type of visualisation for comparison. Histogram plots are aligned. The superpixel sizes, colorbar range and scale are identical between the two examples.

et al. 2010) and anteroposterior and dorsoventral wing disc boundaries (Landsberg et al. 2009; Major and Irvine 2005; Major and Irvine 2006) suggest the importance of physical properties of cell/cell interactions for boundary formation and that collective migration is required for stable boundaries between epithelial populations. To test the importance of physical contact between cells in boundary formation, we cultured the same cell combinations in serum free medium to reduce cell-cell contacts (Fig.5.5A). Under these conditions we observed loss of collective sheet migration (Fig.5.5B,C) as assessed by (normalised) spatial correlation defined as:

$$SCORR(r) = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[\frac{\mathrm{Cov}(V_i V_j)}{\sigma_{V_i}\sigma_{V_j}}\right]_{j\in\mathcal{N}_i}, \mathcal{N}_i := \{j | d(\mathbf{r}_i(t=0), \mathbf{r}_j(t=0)) < r w_s\} \tag{5.3}$$

where $V(t) = \mathbf{r}(t+1) - \mathbf{r}(t)$, $\mathbf{r}(t)$ is the track (all (x, y) positions) up to time t, n is the number of superpixels, $\mathbb{E}[\cdot]$ is the mean function, Cov the covariance function, σ_{V_i} is the standard deviation of $V_i(t)$, r is the distance in multiples of average superpixel width (w_s) and $\mathcal{N}_i$ the neighbouring superpixels of i. t was taken to be the gap closure time (defined later, Ch.5.3.2). No boundary formation was observed and all cell combinations appeared to exhibit similar motion dynamics, (Fig.5.4 0% FBS videos). As further validation of the lack of collective motion in 0% serum, the frame-to-frame correlation between superpixel patches along the produced MOSES superpixel tracks was computed (Fig.5.6). For representative EPC2:CP-A videos it can be seen qualitatively that the extracted temporal image patch sequences show a reduced ability to maintain tracking of individual cells within a superpixel patch due to the uncorrelated independent movement of cells within it in the 0% serum and quantitatively lower average frame-to-frame correlation across all superpixels. This shows that cell-cell contacts are required for boundary formation in our cells. In subsequent experiments we used serum-free conditions as a negative control for boundary formation.

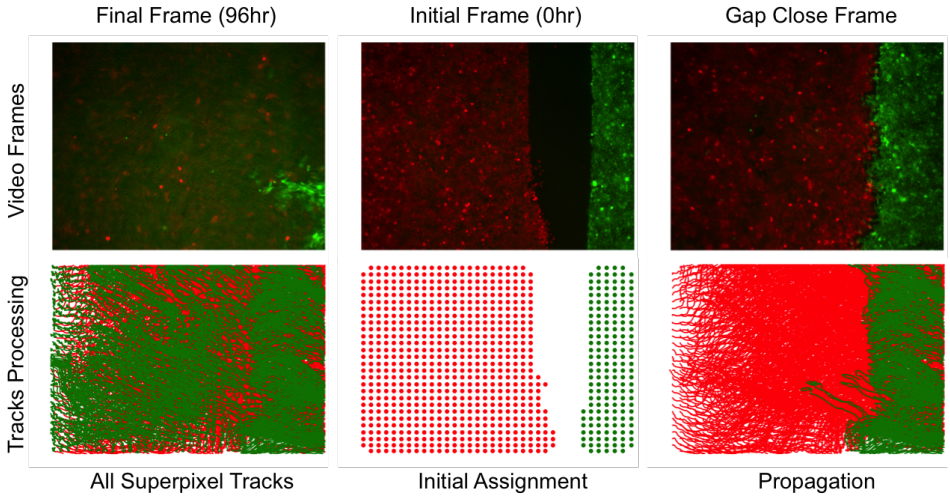

Figure 5.7: Intensity independent superpixel assignment. Top panels show images from an example video at 96hr (left), 0hr (middle) and at the time of gap closure (right). Bottom left shows the 'raw' superpixel tracks plotted from 0-96hr. Bottom middle: shows the assigned red and green superpixels depicted as coloured dots marking their centroid positions at 0hr for the image shown above. Bottom right shows all filtered superpixel tracks from 0-96hr. The extent of the green tracks after assignment now correctly resembles the shape of the interface between the two sheets for the image above.

5.3 Motion Analysis

Motion superpixel tracks were extracted as described in steps 1-3 of the general MOSES framework, (Fig.4.1). Due to spatiotemporal variations in image intensity, automatic assignment of superpixels by image segmentation is not robust. To overcome this, an intensity-independent segmentation method using the spatial layout of the superpixels for migrating epithelial sheets was developed (Fig.5.7).

5.3.1 Intensity Independent Superpixel Assignment

Inputs: Red (R) and green (G) superpixel tracks $\mathcal{S}^R$, $\mathcal{S}^G$ with a total of N superpixels in each:

Step 1: For each colour independently, identify all superpixels that move where 'move' is defined as the subset of superpixels, S_{move} which moves within a preset number of frames denoted t_c, (here we used 2 frames or 2hrs) a cumulative distance greater than 0 pixels where $|\cdot|$ is the euclidean distance.

$$\mathcal{S}_{\text{move}} := \left\{ i \middle| \left\{ \sum_{t=1}^{t_c} |S_i(x_t, y_t) - S_i(x_{t-1}, y_{t-1})| \right\} > 0, \quad i = 1, 2, ..., N \right\}$$

Step 2: Form the neighbourhood adjacency graph, connecting together superpixel i in $\mathcal{S}_{\text{move}}$ to superpixel j in $\mathcal{S}_{\text{move}}$ if their initial distance at $t = 0$, $d\Big(S_i(x_0, y_0), S_j(x_0, y_0)\Big) \leq r_c$ is smaller than a preset distance cutoff (r_c =1.2w_s (average superpixel width) here). The largest connected graph component is found and approximates the initial area covered by each coloured epithelial sheet at frame 0.

Step 3: In some cases, image artifacts such as autofluorescence or the presence of isolated cells detached from the main sheet introduces additional global motion in the optical flow motion extraction resulting in 'noisy' superpixel tracks that may affect statistically significant quantification of motion dynamics downstream. Tracks associated with these noise sources must therefore be removed. Assuming no overlap between red and green superpixels, that is a superpixel at (x, y) cannot both be associated with red and green colours simultaneously at $t = 0$, joint filtering based on which colour exhibits greater movement is applied to clean segmentation errors from steps 1-2. To save computational time, the need for removal is determined by a single user-set cutoff based on prior knowledge of the maximum expected fraction of the field of view covered by any of the red and green populations at $t = 0$ e.g. for a 50-50 plating of red and green cells a reasonable cut-off is 0.7 (used here) corresponding to at worst 70% of the field of view being covered initially by either red or green cells in a high-throughput screening.

Step 4: The kept superpixels after steps 1-3 form an initial set of superpixels $\mathcal{S}^R_{\text{initial}}$, $\mathcal{S}^G_{\text{initial}}$ that cover the initial red and green sheets at time $t = 0$. Ideally these initial superpixel tracks alone would fully recover the full motion dynamics of the respective red and green sheets. In practice however, 'drift' errors occur during long-term tracking due to the accumulation of small errors between successive frames. Thus to improve the coverage of the full dynamic motion in subsequent frames especially at the front of the migrating sheet, superpixels not kept so far or inactivated i.e. $\mathcal{S}^R_{\text{inactivated}} = \mathcal{S}^R - \mathcal{S}^R_{\text{initial}}$ and $\mathcal{S}^G_{\text{inactivated}} =$

$\mathcal{S}^G - \mathcal{S}^G_{\text{initial}}$ is 'activated' and added to $\mathcal{S}^R_{\text{initial}}$ and $\mathcal{S}^G_{\text{initial}}$ respectively if their initial (x, y) position lies within the combined area $\Omega^t := \{\omega^t_1, \omega^t_2, ..., \omega^t_i\}$, the concatenation of the individual areas ω^t_i of superpixel i in $\mathcal{S}^R_{\text{initial}}$, $\mathcal{S}^G_{\text{initial}}$ respectively at time t as defined in Eqn.(4.1). The 'activation' process is carried out iteratively frame-by-frame thus superpixels activated at time t, $\mathcal{S}^t_{\text{activate}}$ is combined with $\mathcal{S}_{\text{initial}}$ to form an augmented set of superpixels for time $t+1$, $\mathcal{S}^{t+1}_{\text{initial}} \leftarrow \mathcal{S}_{\text{initial}} \cup \mathcal{S}^t_{\text{activate}}$. A visual illustration of this step is depicted in Fig.4.1, step 4 for the red sheet where red dots are the centroids of the initial red superpixels, $\mathcal{S}^R_{\text{initial}}$, black dots the centroids of inactivated superpixels, $\mathcal{S}^R_{\text{inactivated}}$ and blue dots the centroids of all initially 'inactivated' superpixels that become 'activated', $\mathcal{S}^t_{\text{activate}}, \forall t$.

Step 5: To ensure the same number of superpixel tracks, N across all videos for statistical comparison, constant tracks is assigned for all inactivated superpixel tracks, $j \in \mathcal{S}_{\text{inactivate}}$ such that for all frames t their (x_t, y_t) positions is fixed to their initial position (x_0, y_0), $S_j(x_t, y_t) = S_j(x_0, y_0)\ \forall t$.

Outputs: Filtered red (R) and green (G) superpixel tracks, $\mathcal{S}^R_{\text{filt}}$, $\mathcal{S}^G_{\text{filt}}$ with a total of N superpixels more specific to the epithelial sheet dynamics and robust to intensity artifacts such as autofluorescence.

Computational Time: for a single 145 frame, two-channel red, green video with a frame resolution of 1344×1024 pixels, tracking 1000 superpixels, steps 1-5 above takes ∼4 mins on a single CPU (3.2GHz, 16GB RAM).

5.3.2 Automatic Gap Closure Determination

The frame in which the gap is closed or the gap closure point is required in order to compare the difference in migration dynamics when the two epithelial sheets are initially separate and following contact. Due to the potential ambiguity in defining the gap closure point for some videos and to facilitate analysis of large video sets, an automatic algorithm to determine the gap closure was developed by finding the

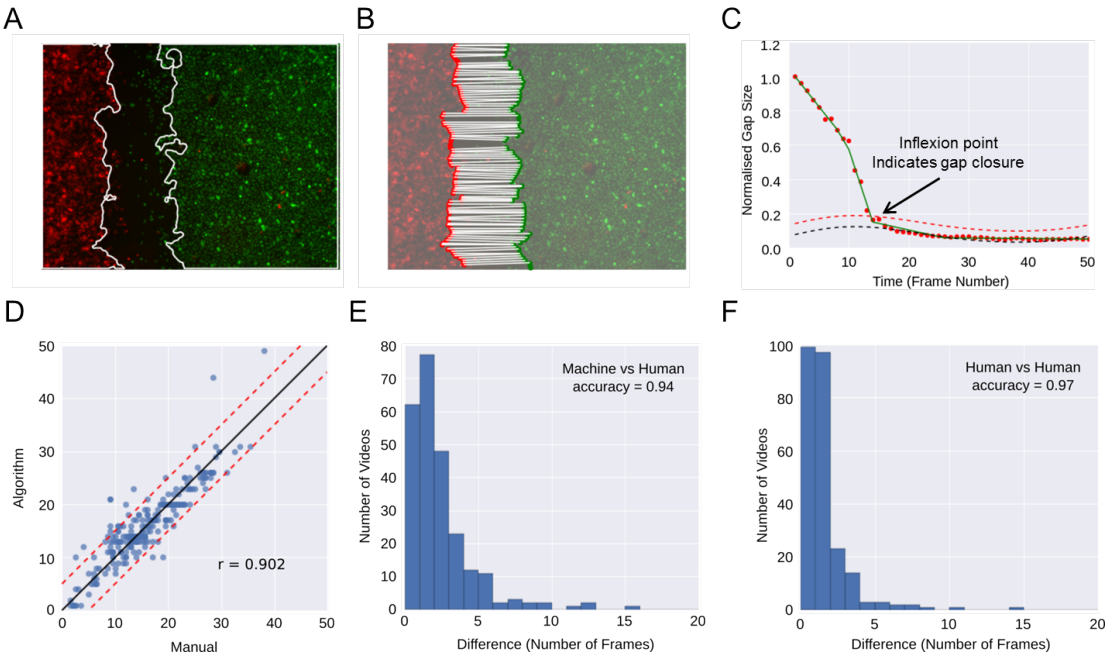

Figure 5.8: Automatic determination of gap closure. A) Epithelial sheet segmentation using the individual red and green image channels for each RGB video frame. Solid white lines mark the sheet boundaries. B) Points identified using a sweepline algorithm subsample the respective coloured sheet boundaries. Each red point is paired with the closest green point, shown with white solid lines. C) The gap closure frame is the point of inflexion in the plot of the average pairwise distance between boundary points (normalised gap size) vs time (given as frame numbers). The inflexion point is automatically estimated by finding the point of intersection between the fitted linear spline approximation (green line) of the distance (red dots) and the fitted baseline + 2 times standard deviation (dashed red line). The baseline (dashed black line) attempts to estimate the minimum separation between the two sheets from the derived distance curve. Ideally this value should be 0 when the two sheets meet, but is not in practice due to errors in image segmentation. D) Comparison of frame number at time of gap closure predicted by the algorithm vs the consensus of two humans given by the average of their independently annotated frame number. Each blue point is a video, n=246 videos. Solid black line is the ideal identity line. Dashed red lines show ±5 frames from the solid black ideal identity line. E) Histogram of the error measured as absolute difference in frame number between automated and consensus human annotations. F) Histogram as in E), showing difference between the two individual human annotators. In E) and F) the accuracy score is reported treating a difference > 5 frames as a disagreement.

video frame that minimises the distance between the boundary points of the two diametrically opposed epithelial layers.

Inputs: Separate red and green channel videos with a maximum number of $T_{\max}$ frames and image size $m \times n$ pixels. $T_{\max}$ is chosen to be long enough to guarantee gap closure but smaller than the maximum number of video frames, T to save computation time.

Step 1: Image Segmentation. Segment each epithelial sheet independently for each frame based on image pixel intensity. For each sheet, preprocess images by applying a median filter with a square kernel of w_s (the average superpixel width) and segment using 2 class K-means clustering for videos with collective motion or 3 class K-means (keeping the top 2 classes with highest image intensities for non-collective motion to capture weaker stained leading cells). The resulting binary images is cleaned by morphological closing (disk kernel of 5) followed by removal of small connected objects (<5% total image area) before binary filling.

Step 2: Locate Boundary Points. A sweepline algorithm is used to locate boundary points efficiently. The image is evenly divided in the y direction into 100 strips or sweeps. For each sweep, one boundary point is identified by selecting the furthest right point if the sheet is moving right or furthest left if the sheet is moving left (Fig.5.8B).

Step 3: Compute Mean Gap Distance Frame-by-Frame. Pair each of the red/green boundary points to the closest in the opposing colour. The average euclidean distance between the boundary points provides an estimate of the mean gap distance between the two sheets at a particular time t. The average gap distance is computed frame-by-frame and normalised to a value between 0 and 1 by dividing by the maximum distance to yield a normalised gap size curve (Fig.5.8C).

Step 4: Locate Inflexion Point. When the gap closes there occurs a sharp change in the rate of decrease of the normalised gap size curve (Fig.5.8C). To estimate the frame where this transition occurs, first the baseline representing the contribution to the gap size curve due to image noise is estimated using the method of asymmetric least means squares (Peng et al. 2010), (black dotted line). To reduce the effect of outliers, a linear spline approximation of the gap size curve is obtained with smoothing factor 0.1Var(curve). The gap closure frame is then the first frame for which the spline curve (solid green curve) falls below the baseline + 2Std(baseline), (red dashed curve) where $\mathrm{Std}, \mathrm{Var}$ is the standard deviation and variance.

Output: The estimated video frame in which the gap between red and green sheets are closed.

The algorithm was validated using a total of n=246 videos of different cell combinations in different media by comparing the automatically inferred frame to the consensus (average frame) of two manual annotators. There was a strong Pearson correlation coefficient of r=0.902 (Fig.5.8D) and a concordance of 94% within ± 5 frames (Fig.5.8E) compared to a concordance of 97% within ± 5 frames between human annotators (Fig.5.8F). For calculating concordance a difference of > 5 frames was considered a disagreement. Almost all the errors were due to videos in no serum where cells show individual movement. The non-coordinated motion of cells in the sheet leads to the presence of isolated populations of leading cells in front of the main sheet. This leads to greater ambiguity in determining the frame of gap closure. Other sources of error were due to weak staining of cells.

5.4 Quantitative measurement of squamous and columnar epithelial boundary formation using MOSES

Using MOSES, quantitative measurements can be readily derived to assess a particular biological feature. For assessing stable boundary formation three indices

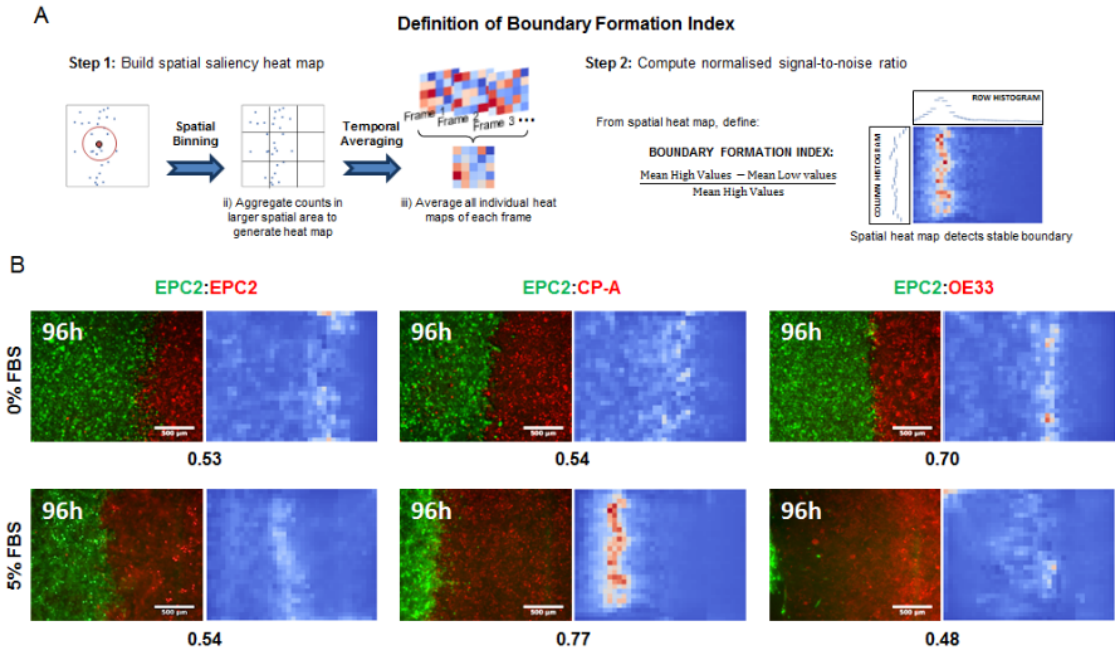

Figure 5.9: Boundary formation index for two epithelial sheets. A) Illustration of the steps in calculating the boundary formation index. B) Example video snapshots at 96 hr of the three cell combinations (EPC2:EPC2, EPC2:CP-A, EPC2:OE33) in 0% and 5% serum with corresponding spatial saliency map and computed boundary formation index below each image pair.

are proposed: i) boundary formation index; ii) motion stability index; and iii) maximum velocity cross-correlation. Boundary formation and motion stability are mesh-derived statistics. Maximum velocity cross-correlation is derived from the individual superpixel tracks. These descriptors/measures are examples of statistics that can be derived using MOSES.

5.4.1 Boundary Formation Index

The boundary formation index is a global measure of the likelihood of stable boundary formation, exploiting the notion that a boundary is a spatially constrained phenomenon. Computing the motion saliency map defined in Ch.4.4.3, (Eqn.(Motion Saliency Map)) integrated over all time using the radius neighbours mesh defined in Ch.4.3.2 for enhanced sensitivity, (Eqn.(Radial Neighbours Mesh)) generates a heatmap that spatially detects the interface between the two migrating sheets from which the global likelihood of boundary formation can be quantified

using a simple normalised signal-to-noise ratio index from 0-1, (Fig.5.9) similar to that used proposed for zebrafish immunosurveillance (Ch.4.6.3); the higher the score the greater the likelihood of boundary formation.

5.4.2 Motion Stability Index

The motion stability index is a local measure that assesses the additional stability of the interface between the two cell populations not captured by the boundary formation index, where individual cells in the sheet may still be moving stochastically without affecting the equilibrium spatial position and shape of the boundary. More precisely, it defines 'stability' as the zero movement state at the superpixel-level when the dynamic mesh no longer deforms and so the relative distance between individual superpixels and its neighbours are fixed. It can be seen as a 'global' stability measure that takes into account collective motion as a superpixel typically encompasses several cells within its area such that the distance between superpixels (and by extension cell groups) may not change and the mesh is stable but the cells within a single superpixel could still be moving and rearranging. Under this definition states such as a constant rate expansion or contraction of the sheet is deemed unstable unless the topology is simultanenously preserved. It is computed by subtracting from 1 the gradient of the normalised mesh strain curve which is the MOSES forward motion signature (Ch.4.4.2, Eqn.(4.12)) divided by its maximum value. Here the normalised mesh strain curve measures the relative degree of motion of superpixels with respect to its neighbours within the tissue. The higher the motion stability index (maximum value of 1), the more stable the boundary and the cells at the boundary. The motion stability index for a cell combination is computed from the averaged normalised mesh strain curve of the individual cell populations. For stable computation of gradients without fitting, for 96 hr videos, the period of stability given by the curve plateau is shorter, therefore the last 10 frames (10 hrs) were used for computing the gradient. For 144 hr videos, the last 24 frames (24 hrs) were used.

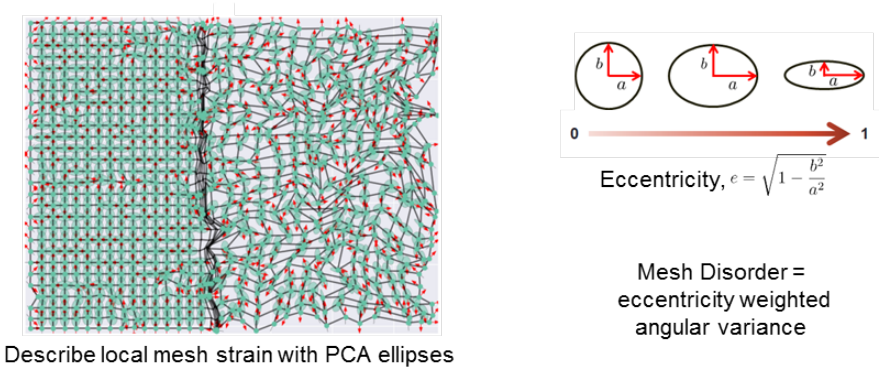

Figure 5.10: Definition of the mesh disorder index as the eccentricity weighted angular variance based on the MOSES mesh. The higher the index the less regular the mesh. Red arrows on the mesh are the computed PCA ellipse orientation.

5.4.3 Maximum Velocity Cross-Correlation

The maximum velocity cross correlation is an averaged normalised measure (0-1) that can measure the degree two epithelial sheets move together as a single cohesive connected sheet by correlating each superpixel track in one population (colour) with every superpixel track of the other population (colour), factoring in time delays (see Ch.3.3.4). For two initially diametrically migrating sheets, a significant increase in this measure before and after closure of the gap is indicative of increased interaction across the sheets. This might be due to increased cell-cell contact in the combined monolayer leading to increased coordinated motion across larger spatial scales. For a video, the maximum velocity cross-correlation is computed for all possible pairings of red and green tracks and the average over all pairings is then reported as evidence of sheet interaction. Based on the accuracy of the automatic gap closure algorithm, the maximum velocity cross-correlation was computed with tracks before (up to -5 frames) and after (from +5 frames) of the inferred gap closure point.

5.4.4 Mesh Disorder Index

The mesh disorder index captures the extent local cell populations move in opposing directions with respect to their neighbours using the spatial irregularity of the mesh. Mathematically this is characterised by the angular distribution of the local strain ellipses, (Fig.5.10). For each superpixel i, $\mathcal{S}_i$, the normalised (by

magnitude) displacement vector $\vec{r}_{ij}(t)$ to each neighbouring superpixel j where 'neighbour' and neighbourhood of i, $\mathcal{N}_i$ is defined by the MOSES mesh, (Ch.4.3.2, Eqn.(MOSES Mesh)) for frame t is determined. PCA is applied on $\{\vec{r}_{ij}(t), j \in \mathcal{N}_i\}$ to solve for the eigenvectors $\vec{r}_{i,1}, \vec{r}_{i,2}$, the respective eigenvalues, λ_1^i, λ_2^i where $(\lambda_1^i > \lambda_2^i)$ and the angle of the principal vector $\vec{r}_{i,1}$, $\theta = \angle\vec{r}_{i,1}$ which is also the principle angle of orientation of the strain ellipse. For an ellipse, the eccentricity, e is defined as $e = \sqrt{(1 - b^2/a^2)}$ where a is the major ellipse length (given by λ_1^i) and b is the minor ellipse length (given by λ_2^i). The mesh order index $MDI(t)$ for frame t is then defined as the eccentricity weighted angular variance, $\mathrm{Var}(e_i \cos\theta_i) + \mathrm{Var}(e_i \sin\theta_i)$ where $\mathrm{Var}(\cdot)$ is the variance operation and measures the variation in the stretch orientation as weighted by the strength of the stretch over all superpixels i. The contribution is smaller for circular ellipses whilst stronger for directional ellipses, (Fig.5.10). Superpixel points that have a large number of surrounding superpixels are 'defect centres' that should be removed since we wish to measure the disorder in the bulk of the sheet not at concentration points such as the boundary and edges for which the MOSES mesh is naturally more compressed, (Fig.5.10). One effective method is to exclude superpixels with a number of neighbours, $n_{\text{neighbours}}$ less than a maximum specified number of neighbours $n_{\max}$ where the neighbourhood of superpixel i is all superpixels j whose (x, y) position at time t is less than a cut-off distance, $r_c = 1.2w_s$ away and w_s is the average superpixel width. For the MOSES mesh, the value of this index strictly is only applicable in describing the collective behaviour of the initial frames prior to gap closure. A single MOSES mesh disorder index for comparison can then be reported as the mean of the MOSES mesh disorder index before gap closure after smoothing the values with a linear spline fit. For motion after closure, the notion of collectiveness in a migrating sheet becomes somewhat ill-defined. The MOSES mesh is disordered due to 'collision' with the other sheet and in most combinations the individual sheets stop moving shortly afterwards. The proposed mesh definition of collectiveness is not 'universal' unlike a physical quantity such as speed. It depends on the initial mesh geometry and therefore should only be used strictly speaking to compare the same cell

type with the same or similar starting conditions as in our EPC2:CP-A EGF titration below (Ch.5.7). In the situation where it might be desired to compare the motion of one cell type in sparse conditions and the same cell type under confluent conditions, the mesh disorder index no longer has clear interpretation. In these two cases they are not influencing the same number of superpixel points and the resultant index will be skewed since the mesh captures not only collectiveness but also differences in speed, shape etc.

5.4.5 Biological Interpretation of Proposed Measurements

Table.5.1 summarises the proposed four indices above and suggests possible biological examples and interpretations. In general, each index is not specific to any particular biological application. Instead, the interpretation depends on the specific cellular motion and table.5.1 should serve only as a guide.

5.5 The Video Dataset

In total 190 videos were acquired for analysis, 125 videos of 'normal' experiments across 4 experiments with no external stimuli and 65 videos of titration experiments where increasing amounts of EGF (epidermal growth factor) is added. Each video was acquired one frame per hour on a Nikon time-lapse microscope. These videos are highly heterogeneous and acquired under different conditions; uneven fluorescence, plated either side (left or right, asymmetrically) and imaged at different magnifications, altogether creating a challenging dataset overall for analysis. An image panel of 35 randomly selected videos from the set of 125 normal videos is displayed showing the video snapshots at 0h, gap closure and 96 (Fig.5.11).

5.6 Squamous-Columnar Cell Combinations Can Form Boundaries

The proposed indices were computed for different combinations of the three cell lines (EPC2, CP-A and OE33) in the experimental system. 125 videos (48 with

Index	Definition	Physical Interpretation	Biological Interpretation	Biological Applications
Boundary Formation Index (value from 0-1)	Signal-to-noise ratio in motion saliency map	Average strength of all motion sinks/sources in video	probability of a particular spatial region attracting / repelling local cellular 'motion', the larger the index the higher the likelihood	locating wound sites for neutrophils and macrophages, locating line boundaries where two epithelial sheets coalesce.
Motion Stability Index (value from $-\infty$ to 1)	Rate of change in the average distance of a superpixel to its neighbouring superpixels	extent of deviation from a global collective state	cell groups are not moving independently with respect to their neighbours but maintaining the same relative spatial arrangement	a static epithelial sheet is stable (scores 1), neighbour rearrangements are unstable (<1), moving cells that preserve topology and distances with respect to its neighbours is stable (scores 1)
Maximum Velocity Cross-Correlation (value from 0-1)	Average correlation between two objects' movement trajectories based on directionality factoring in time delays	0-no correction between moving entities, 1-both entities are perfectly correlated (positive/negative) up to a time delay	presence of motion similarity may indicate interaction between two or more entities through signalling mechanisms, physical binding or social cues	detection of leader/follower relationships between individual cells/birds (Shishika et al. 2014) and detecting coordinated motion
Mesh Disorder Index (Value from 0-1)	Average local angular distortion	quantifies global mesh deformation through the angular orientation of each superpixel to its neighbours	deviation of a cell or cell group from the movement directionality of 'neighbours', due to e.g local rearrangement or loss of collective motion	detection of local deviation from collective motion, detection of cellular rearrangement events, detection of changes in social groups

Table 5.1: Summary and interpretation of proposed measurements for two cell populations

0% serum and 77 with 5% serum) were collected from 4 independent experiments with no external stimuli and jointly analysed, (Table.5.2). Cells grown in 0% serum were used as negative control to set a cut-off for boundary formation (0.69) defined statistically as one standard deviation higher than the pooled mean of all three combinations (Fig.5.12A). Above this cut-off, cell combinations are categorised

		Expt 1 (n=37)	Expt 2 (n=50)	Expt 3 (n=21)	Expt 4 (n=17)
Medium	Cell Combinations	4x, 10μM dye, asymmetric	4x, 10μM dye, symmetric	4x, 10μM, symmetric	2x, 2.5μM, symmetric
5% serum	EPC2:EPC2	3	5	7	2
	EPC2:CP-A	5	8	13	4
	EPC2:OE33	5	7	1	4
	EPC2:AGS	-	-	-	2
	CP-A:CP-A	-	-	-	2
	CP-A:OE33	6	-	-	-
	OE33:OE33	-	-	-	3
0% serum	EPC2:EPC2	6	10	-	-
	EPC2:CP-A	6	10	-	-
	EPC2:OE33	6	10	-	-

Table 5.2: Summary of video datasets and experiments analysed for 0% and 5% serum.

All imaged 2x, 2.5μM dye, symmetric divider, total = 65 videos						
Experiment No.	EGF / Medium	0 ng/ml	2.5 ng/ml	5 ng/ml	10 ng/ml	20 ng/ml
Exp 1	5% serum	-	-	2	2	1
	0% serum	2	-	2	1	-
Exp 2	5% serum	4	4	4	4	4
Exp 3	5% serum	4	-	4	3	4
Exp 4	0% serum	4	4	4	4	4

Table 5.3: Summary of video datasets and experiments analysed for EGF addition.

as forming a boundary. The boundary formation index was highest (0.74) for EPC2:CP-A grown in 5% serum (n=30/77) (Fig.5.12A,D). For EPC2:EPC2 and EPC2:OE33 in 5% serum the boundary formation index was below the cut-off (Fig.5.12A). The serum videos were also ranked on a continuous scale using the boundary formation index to show unbiasedly that the majority of EPC2:CP-A videos are at the top of this ranking (Fig.5.13). Similarly, experiments in 0% serum were used to set the global motion stability threshold (0.87), one standard deviation below the pooled mean of EPC2:EPC2 and EPC2:CP-A. Cells at the interface of EPC2:OE33 are not stable and therefore not included in the pooled statistics. Below this cut-off, cell combinations are categorised as forming unstable interfaces. In 5%

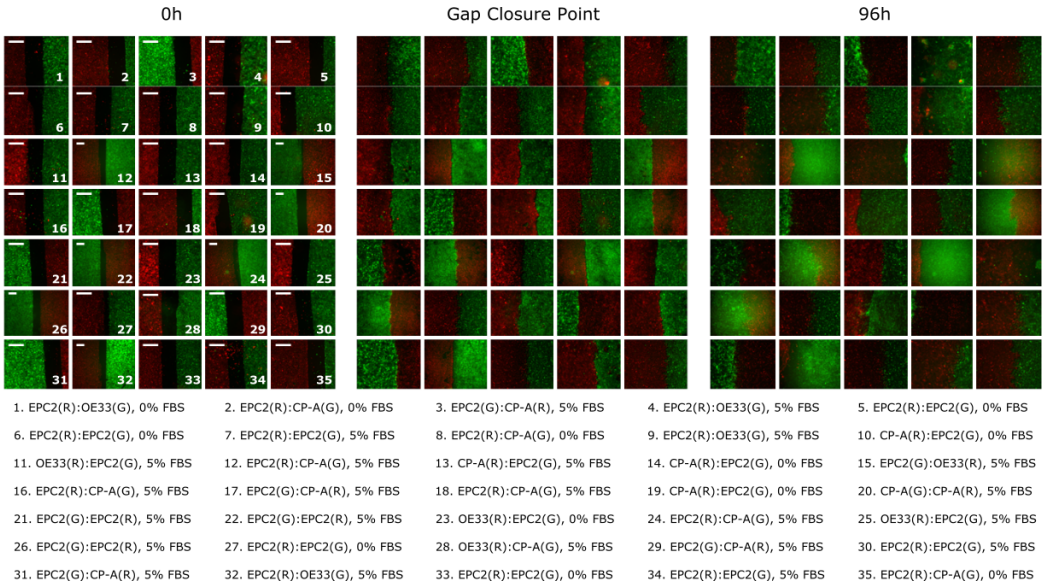

Figure 5.11: Heterogeneity in motion dynamics and quality of image acquisition. A random selection of video snapshots are shown as examples of the video variability. Videos are labelled by number and each cell combination is named such that the first cell name corresponds to the left sheet and the second name to the right sheet. The dye used is indicated as (R) for red and (G) for green. Left: Examples at 0h at the start of the experiment. Middle: Examples when the red and green sheets first close the gap between them. The gap closure point varies among videos and is detected automatically to ± 5 frames with 94% accuracy (Ch.5.3.2). Right: at 96hr. All scale bars: 500μm.

serum, EPC2:CP-A had the highest motion stability index with a median value of 0.90, compared to EPC2:EPC2 (0.76) and EPC2:OE33 (0.25) (Fig.5.12B,E). These results illustrate that the squamous-columnar combination EPC2:CP-A forms a stable boundary.

Sheet-sheet interactions were measured using the maximum velocity cross correlation both before and after gap closure for the three cell type combinations. In 0% serum there was no difference in velocity cross correlation across all combinations before and after gap closure (Fig.5.12C). The two sheets do not move cohesively as an unit, which is to be expected with minimal cell-cell contact. In serum, there is greater cohesion due to cell-cell contact. We observe that the difference for EPC2:CP-A (median value of 0.03 before and 0.20 after gap closure) was ∼3-6 times larger than for EPC2:EPC2 (0.01 to 0.08) and EPC2:OE33 (0.02 to 0.05), (c.f. left and right violins in Fig.5.12C). We note also that CP-A:OE33 (Barrett's:cancer,

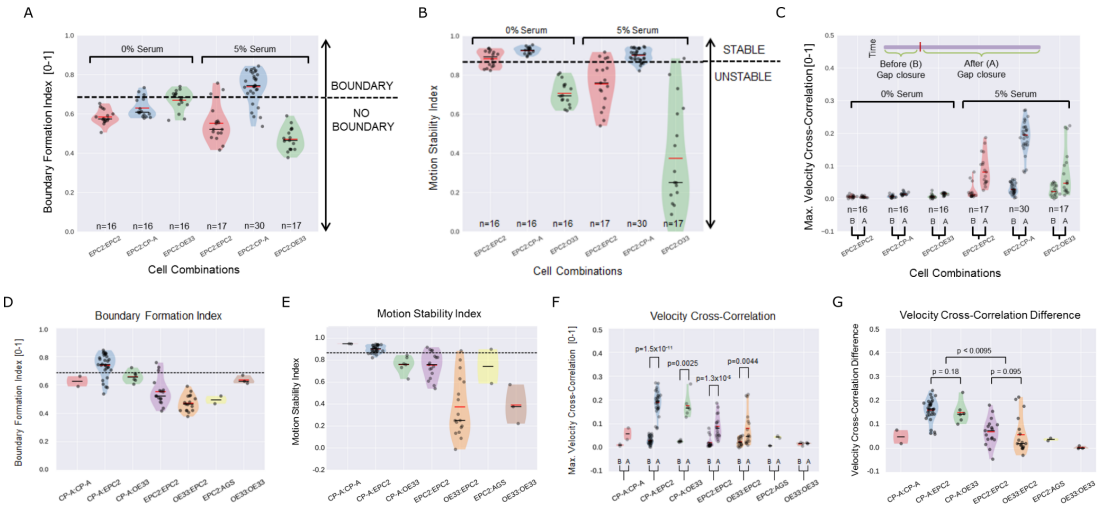

Figure 5.12: Quantitative assessment of boundary formation and sheet-sheet interaction dynamics. A) Violin plots of boundary formation index for each video (black dot) for the three main cell combinations in 0% and 5% serum. Dashed line is the threshold given by one standard deviation above the pooled mean value of all cell combinations in 0% serum. Red solid line = mean, black solid line = median. Shaded region is the probability density of the data whose width is proportional to the number of videos at this value. B) Motion stability index and C) Maximum (Max.) velocity cross-correlation between the two sheets. 'B' is before and 'A' is after gap closure. Violin plots of D) Boundary formation, E) motion stability and F) maximum (max.) velocity cross-correlation between the two sheets for all cell combinations tested in 5% serum. In panel F), 'B' is before and 'A' is after gap closure. G) Velocity cross-correlation difference before and after gap closure of F). Statistical two-tail test at 5% significance level with Mann-Whitney U. For all panels, each black point is a video, red solid line is the mean, black solid line is the median, shaded region is the probability density of the data whose width is proportional to the number of videos at this value, n=77.

n=6) also exhibited a substantial increase in velocity cross-correlation following gap closure (0.03 to 0.17) (Fig.5.12F,G). This is unlikely to be a feature of the CP-A cell line, as no substantial increase was observed for CP-A:CP-A (0.01 to 0.06) (Fig.5.12F,G). Thus, the EPC2:CP-A boundary exhibits greater cohesion between the two cell populations compared to interfaces formed between cells of the same type and EPC2:OE33 (i.e. 'normal' squamous cells with cancer cells) suggesting enhanced physical interaction. Of all the cell combinations tested, the squamous-columnar combination EPC2:CP-A uniquely forms an 'interacting' stable boundary.

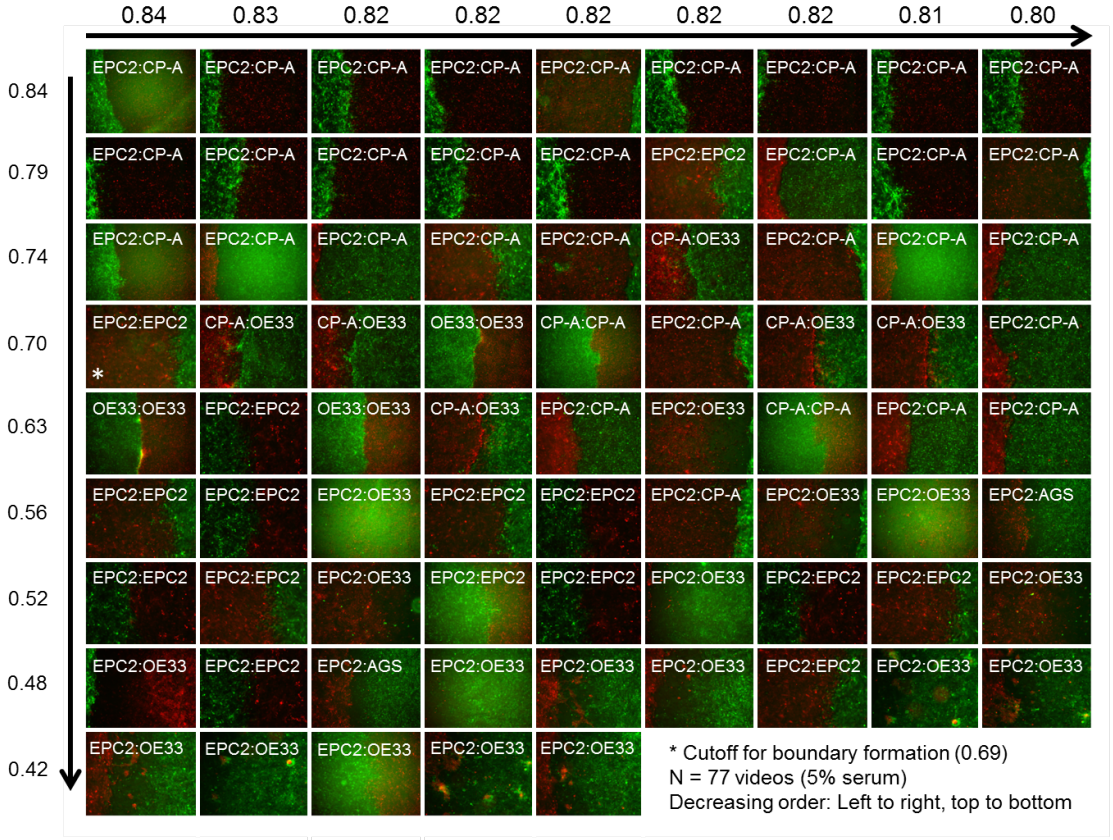

Figure 5.13: Ranking of 5% serum videos according to boundary formation index. Final frame snapshots (96 or 144 hrs) ordered according to the MOSES boundary formation index, decreasing order reading from left to right, top to bottom, (n=77). Values to 2.s.f (significant figures) are given for the first video in each row and column. * demarcates the first image where a boundary does not form, according to the threshold set by our negative 0% serum controls from Fig.5.12A (0.69).

5.7 Measuring Subtle Phenotype Changes Induced by External Stimuli

To be relevant to high-content imaging analyses, MOSES needs to quantitatively monitor subtle changes in cellular motion dynamics with a minimal number of replicates. As a test of the sensitivity of MOSES, we assessed whether MOSES could detect subtle changes in the EPC2:CP-A boundary caused by an external stimulus.

The main cause of BE (Barrett's Esophagus) is bile acid reflux, (Dixon et al. 2001; Souza 2010). Bile acid activates epidermal growth factor receptor (EGFR)

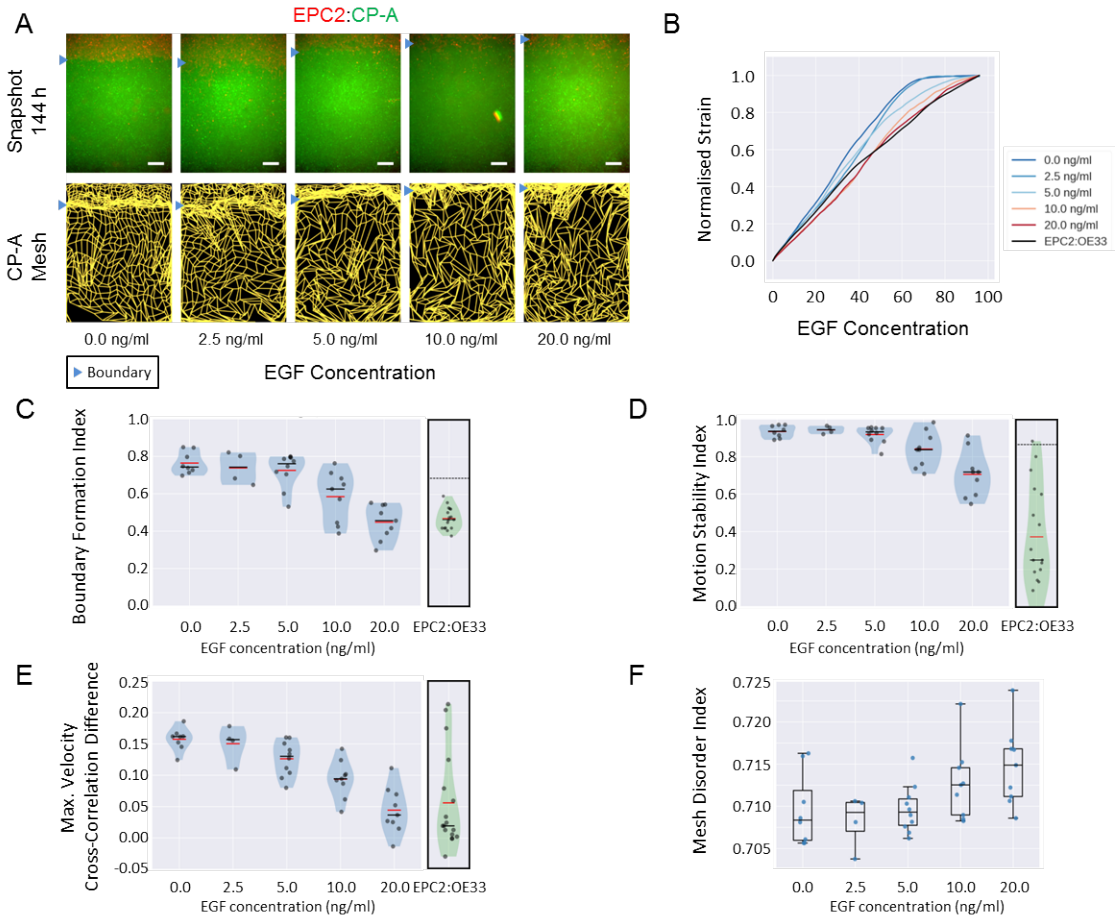

Figure 5.14: EGF titration at physiological levels disrupts boundary formation. A) Destabilisation of the junction with EGF addition. All in 5% serum, snapshots at 144hrs of the green CP-A MOSES mesh at the same timepoint for all concentrations. The closeness of the lines indicates impeded motion leading to a local aggregation of superpixels in the vicinity and suggestive of a boundary. The less rectangular the mesh, the less ordered the motion. Blue triangles mark the boundary position in the image and its corresponding inferred position in the CP-A mesh. All scale bars: 500μm. B) Mean normalised strain curves for each concentration. The mean curve for EPC2:OE33 videos in 5% serum without EGF in Fig.5.12 is shown for comparison. Violin plots of boundary formation index C), motion stability index D) and maximum velocity cross-correlation E) for each concentration of EGF. Red solid line = mean, Black solid line = median. Dots are individual videos, n=40. Values for EPC2:OE33 in 5% serum without EGF and threshold from Fig.5.12 are shown for comparison. Shaded region is the probability density of the data whose width is proportional to the number of videos at this value. F) Boxplot of mesh disorder index for each concentration of EGF.

(Dossa et al. 2015), a receptor tyrosine kinase that is frequently mutated in EAC, (Secrier et al. 2016) and sometimes overexpressed in BE, (Al-Kasspooles et al. 1993; Cronin et al. 2011). We therefore used EGF, the ligand of EGFR, as a stimulus to activate the EGFR signalling pathway. Increasing amounts of EGF (0ng/ml to 20ng/ml) were added to the culture medium to assess incremental effects on cellular motion and boundary formation in the EPC2:CP-A combination. Viewing the mesh (Fig.5.14A), shows the boundary position further away from the initial point of contact between the two cell populations and decreased coherence of the boundary with increasing EGF. This is also shown by the shape of the mean normalised strain curve (Fig.5.14B): at 0ng/ml EGF this curve linearly increases before plateauing around 72hrs; as EGF concentration increases, the curve becomes more linear and the plateau is lost above 5ng/ml. The boundary formation index decreases with increasing EGF (0.74 at 0ng/ml to 0.46 at 20ng/ml), indicating that the boundary is lost (i.e. index below the 0.69 cut-off) (Fig.5.14C). The index at 20ng/ml EGF is similar to that for EPC2:OE33 without EGF (0.46), (Fig.5.14C). The interface becomes increasingly unstable and cells move more as the motion stability index decreases from 0ng/ml (0.94, stable) to 20ng/ml (0.72, unstable) (Fig.5.14D). Also, the interaction between the two cell populations is lost as the maximum velocity cross correlation difference before and after gap closure decreased from 0ng/ml (0.16) to 20ng/ml (0.04) EGF (Fig.5.14E). The maximum velocity cross correlation after gap closure is similar to that for EPC2:OE33 (0.02) (Fig.5.14C), but the motion stability index remains higher (Fig.5.14D). Altogether these measures show that above 5ng/ml EGF the phenotype of EPC2:CP-A becomes similar to that of the interaction between EPC2 and the EAC cell line OE33.

The mesh disorder index showed statistically significant increases with EGF concentration (0ng/ml: 0.708, 2.5ng/ml: 0.709, 5ng/ml: 0.709, 10ng/ml: 0.713, 20ng/ml: 0.715) (Fig.5.14F, 5.15A,B), suggesting collective sheet motion is lost. This loss of collective motion is supported by a decrease in the standard spatial correlation measure (Ch.5.2.3). However, using this standard approach gives high

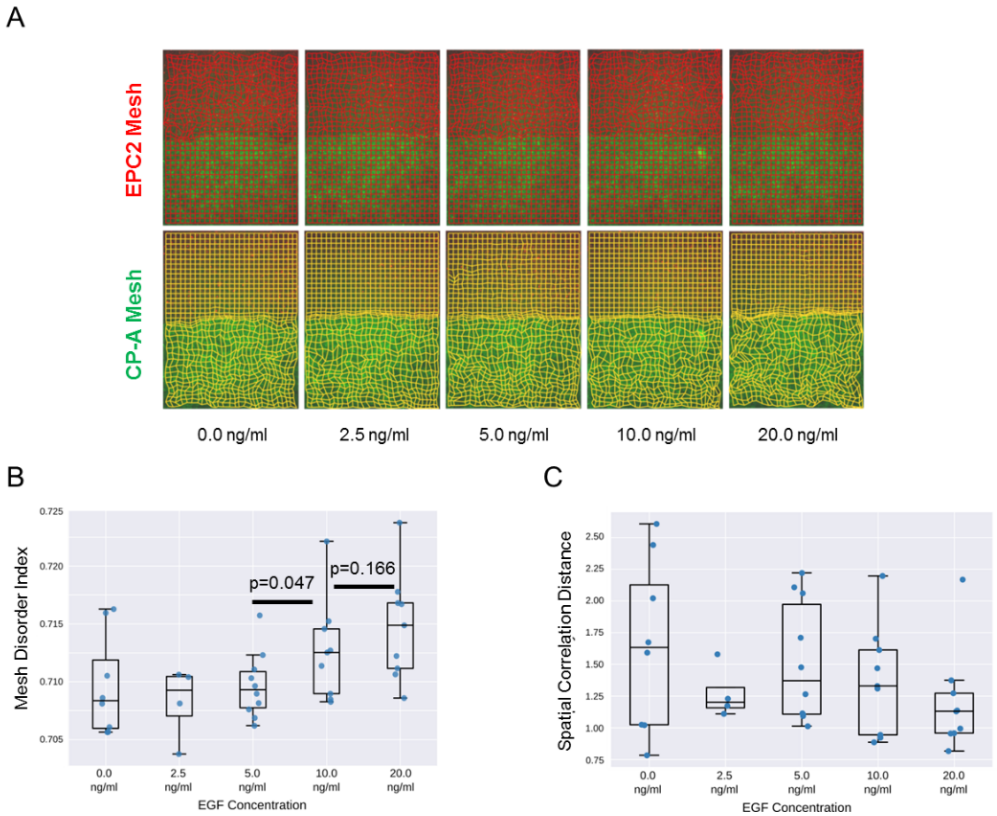

Figure 5.15: Increased mesh disorder with EGF addition to EPC2:CP-A in 5% serum. A) Snapshot of the EPC2 and CP-A meshes at the point of gap closure for different EGF concentrations. Box plots of MOSES mesh disorder index B) and spatial correlation C), black centreline is the median. Each video is a point in C) and D), n=40. Statistical two-tailed test with Mann-Whitney U. No statistically significant differences were detected using spatial correlation between any pairing in D) at 5% significance level.

statistical variance, (Fig.5.15C). The effect of increased disorder is visually subtle, but is clearly detected using the MOSES mesh and the proposed mesh disorder index.

Titrating EGF in the absence of serum gave non-significant changes in the boundary formation index (0ng/ml: 0.60±0.07, 20ng/ml: 0.65±0.03) and maximum velocity cross correlation (Fig.5.16 (n=25)). Decreasing motion stability index indicates increased cell movement without collective motion leading to more distorted meshes (Fig.5.16A) as evidenced by an increasing mesh disorder index . In summary, this example with EGF in the context of our experimental set-up shows that MOSES enables continuous-scale quantification of motion after systematic perturbation in a medium-throughput 24-well format.

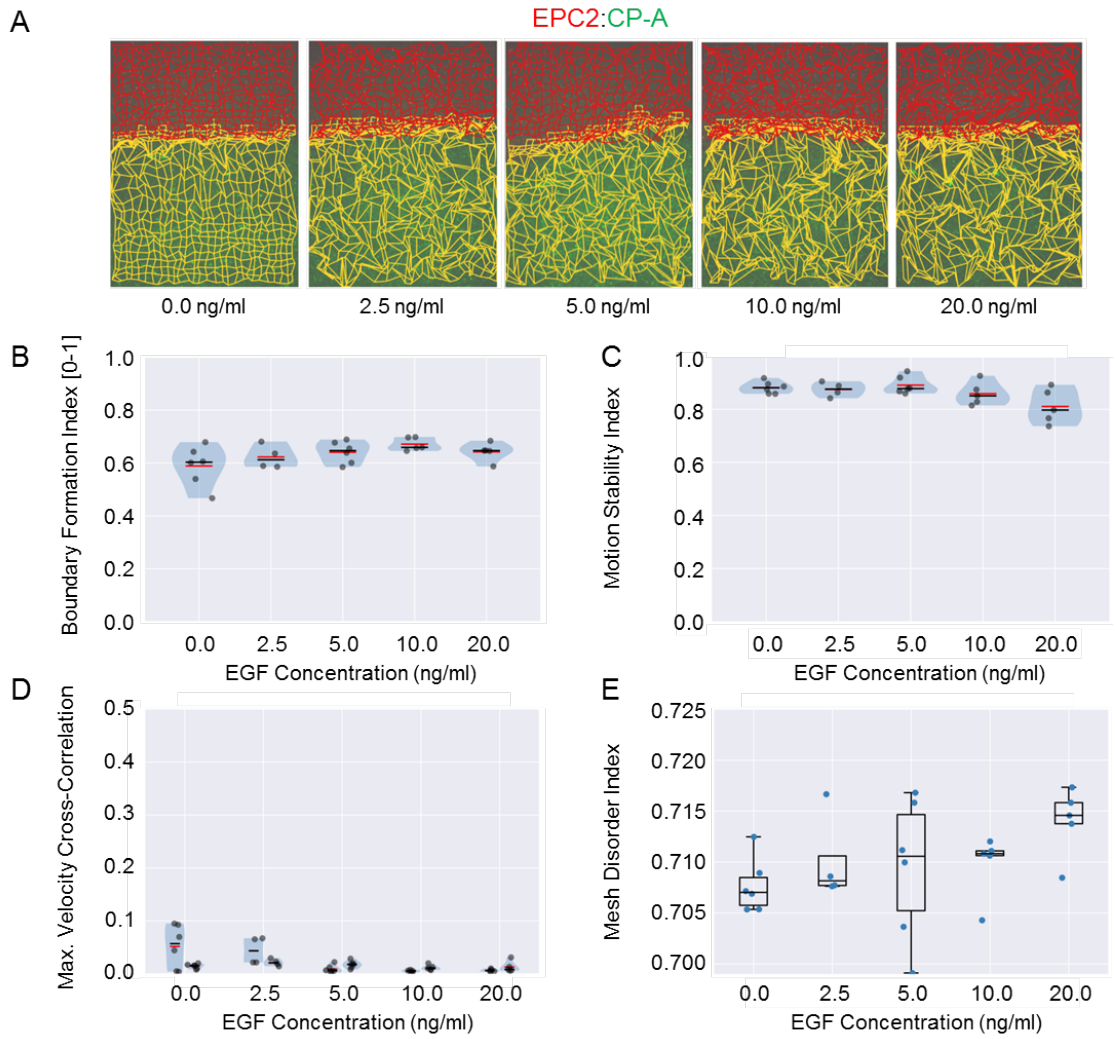

Figure 5.16: EGF addition to EPC2:CP-A in 0% serum does not induce boundary formation. A) Snapshots of the merged red EPC2 and green CP-A MOSES mesh at 144hr with increasing concentration. Violin plots of B) boundary formation index, C) motion stability index and D) maximum (max.) velocity cross-correlation before (B) and after (A) gap closure. Shaded region is the probability density of the data. Width of shaded region is proportional to the number of videos with that value. E) Box plots of mesh disorder index. For B)-E) each point is a video, black line is the median, red line is the mean.

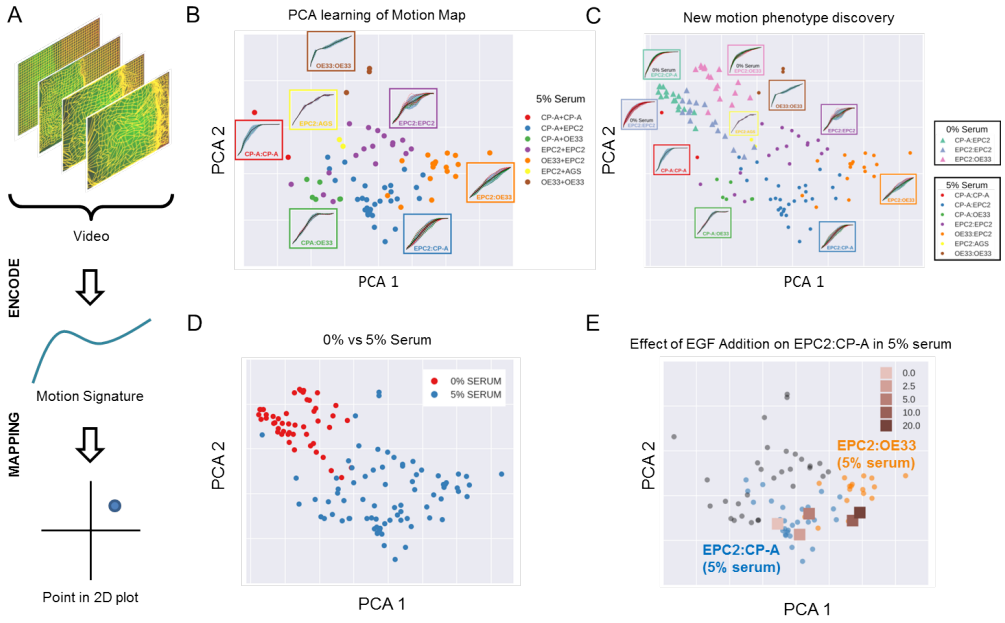

Figure 5.17: 2D motion map for unbiased characterisation of cellular motion phenotypes. In all panels, each point represents a video (see legends for colour code). The position of each video on the 2D plot is based on the normalised mesh strain curves, analysed by PCA. A) The mapping process for a single video. B) The 5% serum videos (n=77) were used to set the PCA that maps a strain curve to a point in the 2D motion map. C) The 0% serum videos (n=48) was plotted onto the same map defined by the 5% serum videos using the learnt PCA. In B) and C) the mean mesh strain curves for each cell combination are shown in insets. Light blue marks the two standard deviations with respect to the mean curve (solid black line). D) Same map as in C) with points coloured according to 0% or 5% serum. E) The normalised mean strain curves for 0-20ng/ml EGF addition to EPC2:CP-A from Fig.5.14 plotted onto the same map defined by the 5% serum videos.

5.8 Motion Signatures and 2D Motion Maps for Unbiased Characterisation of Cellular Motion Phenotypes

The MOSES mesh-based formulation can thus derive robust measurements to test directly for differences in expected motion phenotypes, such as boundary formation. But what if we do not know *a priori* what motion phenotype to expect? For example, in high-content screens it is necessary to identify unknown differences in complex cellular motions unbiasedly from a large number of videos. MOSES addresses this need by facilitating the systematic generation of unique 'motion signatures' for each video. Here unsupervised machine learning techniques requiring no manual user

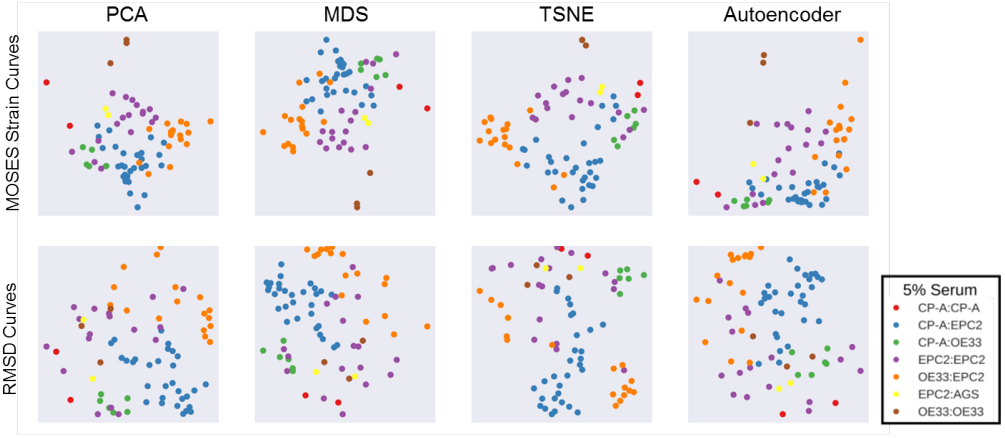

Figure 5.18: Comparison of motion map learning using different dimensional reduction techniques with MOSES strain curves and RMSD curves. In each panel the same 77 serum videos were used. From left to right, PCA - principal components analysis, MDS - multidimensional scaling, TSNE - t-distributed stochastic neighbour embedding and a neural network autoencoder. Each point represents a video, as indicated on the legend.

annotation can be used to plot migration videos onto a 2D motion phenotype map, enabling easy visual assessment of motion phenotype and hypothesis generation without the need to individually interrogate each video.

The general process for motion map generation is illustrated in Fig.5.17A. To position each video on a 2D map, principal components analysis (PCA) is applied to the normalised mesh strain curves used here as a 1D motion signature to summarise the entire video motion for the 77 videos of cell combinations cultured in 5% serum conditions. The generated map (Fig.5.17B) shows that this unbiased approach has automatically clustered together the videos for each cell type combination. Furthermore, the videos are ordered in a continuous manner, as shown by the progressive transformation in the shape of the mean normalised strain curve when looking across the plot in Fig.5.17B from left to right, CP-A:CP-A to EPC2:OE33 (i.e. the shape is increasingly linear). This result could not be achieved with RMSD (root mean squared displacement, c.f. Fig.5.18, 5.19), and is independent of the particular dimensionality reduction technique used, (Fig.5.18). Further, the 1D motion signatures derived from MOSES could be used to train a machine learning classifier with no further processing to predict cell combination identity

better than RMSD, (Fig.5.19). Whilst RMSD and MOSES exhibit similar average scores, looking at the three combinations with the most number of acquired videos (EPC2:EPC2, EPC2:CP-A and EPC2:OE33) MOSES ensures significantly better worst performance of the two (F1 scores ≥ 0.67 vs F1 scores ≥ 0.50). For biological applications arguably even performance across all 'classes' is more important than the average performance which can be dominated by a small fraction of high performing classes to minimise the number of false positives in unbiased screens.

To demonstrate how 2D motion phenotype maps generated by MOSES can be used to compare videos, we next mapped the 48 videos from 0% serum cultures onto the same axes as the videos from 5% serum (Fig.5.17C,D). The videos from 0% serum mapped to a different area of the 2D plot, whilst preserving the continuous ordering of the previous videos. Therefore, without having watched the videos it is easy to predict that the cells have markedly different motion dynamics in 0% serum compared to 5% serum. Furthermore, since the points for the 5% serum videos cover a larger area of the 2D plot than the 0% serum videos, one would expect less diversity of motion in 0% serum, (Fig.5.17D).

The motion map also capture subtle motion changes. This is demonstrated by mapping the mean video motion for each concentration of EGF from 0-20ng/ml (represented by the respective mean normalised strain curves for each concentration (1 per concentration from total n=40 videos, see Fig.5.14B)) onto the same axis as the 5% serum videos in the absence of EGF (square points in Fig.5.17E). With increasing EGF, the EPC2:CP-A motion dynamics becomes more similar to EPC2:OE33 above 5ng/ml, as evidenced by the square points moving from the area of blue circular EPC2:CP-A points into the area of orange circular EPC2:OE33 points. Therefore, the motion map is consistent with the results obtained using the specific derived indices (Ch.5.7). These results illustrate that MOSES is able to account for biological and technical variability unbiasedly across independent experiments and possesses the required features for an algorithm to be used in an

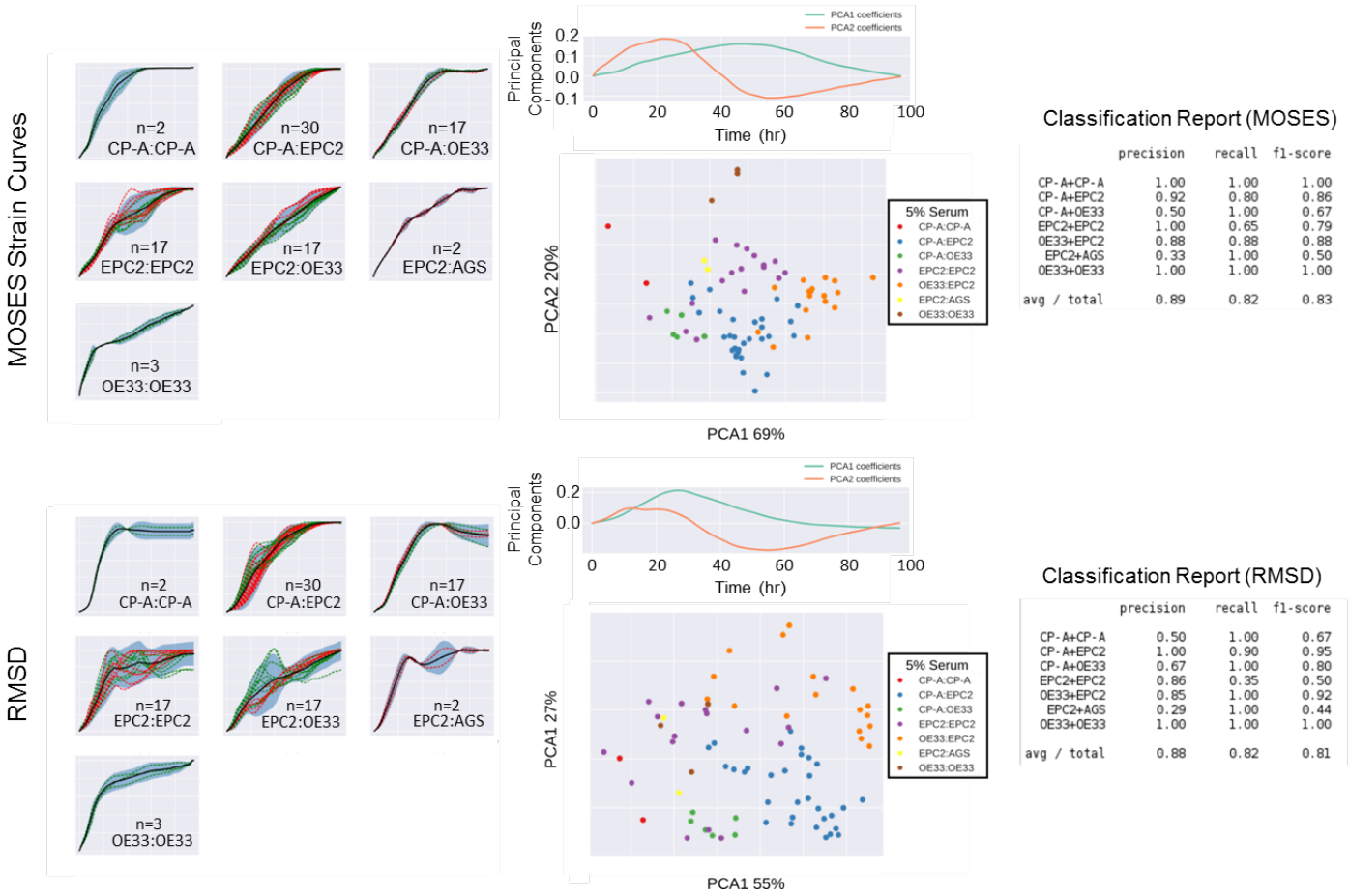

Figure 5.19: Comparison of MOSES normalised strain curves vs RMSD curves as motion signatures for motion map generation. 5% serum videos are used. Left: MOSES normalised strain curves organised by cell combination (top) and RMSD (bottom). Each video is represented by a single dashed curve coloured according to the dye colour used for the first cell type in the combination indicated. Thus for combinations such as CP-A:CP-A where the green dyed CP-A was always the left sheet, there is only one green dashed line. Solid black line is the mean curve, blue shaded region marks 2 standard deviations either side of the mean curve, total n=77 videos. Middle: PCA analysis applied to MOSES (top) and RMSD (bottom) with the respective principal components. Each point is a video (see legends for colour code). Right: Classification report giving the precision, recall and f1 score after fitting a balanced linear SVM (support vector machine) classifier trained on the normalised strain curves of all the videos, n=77.

unbiased manner in high-content screening. For a 96-well plate assay, one 1344×1024 pixel video per well the analysis can be completed in under a day.

5.9 Summary and Conclusions

- Metaplasia at the human oesophageal squamous-columnar junction (Barett's oesophagus) is the most significant risk factor for progression to esophageal adenocarcinoma.
- MOSES is applied to assess the interaction between epithelial cell lines relevant to the oesophageal squamous-columnar junction: EPC2:EPC2 (normal squamous control), EPC2:CP-A (squamous:columnar, Barrett's esophagus) and EPC2:OE33 (squamous:cancer).
- MOSES enables derivation of robust biological metrics for the assessment of boundary formation between different combinations of epithelial cell lines. Three metrics are proposed, i) boundary formation index, ii) motion stability index and iii) maximum velocity cross-correlation.
    1. Facilitates objective assessment of boundary formation in videos from a wide variety of experimental conditions and of low-resolution.
    2. Subtle changes to boundary formation from EGF addition to EPC2:CP-A is captured on a continuous scale.
- A fourth metric, mesh disorder index is proposed as a geometric measure for quantifying changes in collective motion after perturbation for the same cell type with the same starting configuration.
- EPC2:CP-A (squamous-columnar) combination is found to form a unique stable boundary which is disrupted in serum by increasing addition of EGF.
- MOSES signatures facilitates motion map learning for visual comparison of video motion similarity. Resultant motion map enables visual distinction of new motion phenotypes. Continuous phenotype changes can also be monitored for phenotype prediction.
- MOSES provides a simple consistent framework for high-content screening and analysis for 2D *in-vitro* individual and collective migration and interaction.

6
Organoids

Contents

In this chapter the Motion Sensing Superpixels (MOSES) framework is used to automatically monitor organoid growth dynamics filmed with label-free microscopy even with multiple organoids in the field of view. All experiments and imaging were performed by Xiao Qin. Manual organoid counting was carried out by Xiao Qin and Carlos Ruiz-Puig. Manual organoid branch annotations were created

by Carlos Ruiz-Puig.

6.1 Organoids as a Screening Platform

A number of recent research efforts demonstrate the viability of establishing high-throughput screening with organoids. In the first proof-of-concept organoid screen, van de Wetering et al. 2015 use 19 colorectal tumour organoids (derived from 18 different patients) in experimental triplicate to screen the drug sensitivity of a bespoke 83-compound library using organoid viability from IC50 scores. They show using RNA-seq, genetic correlations between individual oncogenic mutations and drug response. Boehnke et al. 2016 further scaled this system to demonstrate the experimental feasibility of an automated 384-well assay format screening with 16 compounds. They proposed a test for spatial uniformity based on the heterogeneity of the response (IC50) of each well to the application of a reference compound chosen as a positive control. Huang et al. 2015 develop ductal pancreatic organoids and demonstrate with 5 organoids and epigenetic regulators that the tumor organoids were capable of retaining patient-specific traits, including repressive epigenetic marks, oxygen consumption and EZH2 dependence. Despite such successes, there still remains technical challenges before organoids can be used for high-content screening and analysis. One major challenge is the biological variability during growth. Current techniques are limited in their precise control of the resulting organoid phenotype for example the spatial differentiation of cell types, size, shape and quantities in each well even if they are expanded from the same primary material. Currently there is a lack of quantitative tools for detailed analysis of long-time imaging of organoid development to facilitate the systematic study of organoid phenotype heterogeneity. Existing methods primarily only use cell viability (dead or alive) to screen organoids for further sequencing analysis or only use area to assess growth dynamics. Both of these approaches neglect the diverse organoid morphologies that are hallmarks of tissue of origin (Fatehullah et al. 2016), (Fig.1.1) that may be perturbed under pathological conditions. For example van de Wetering

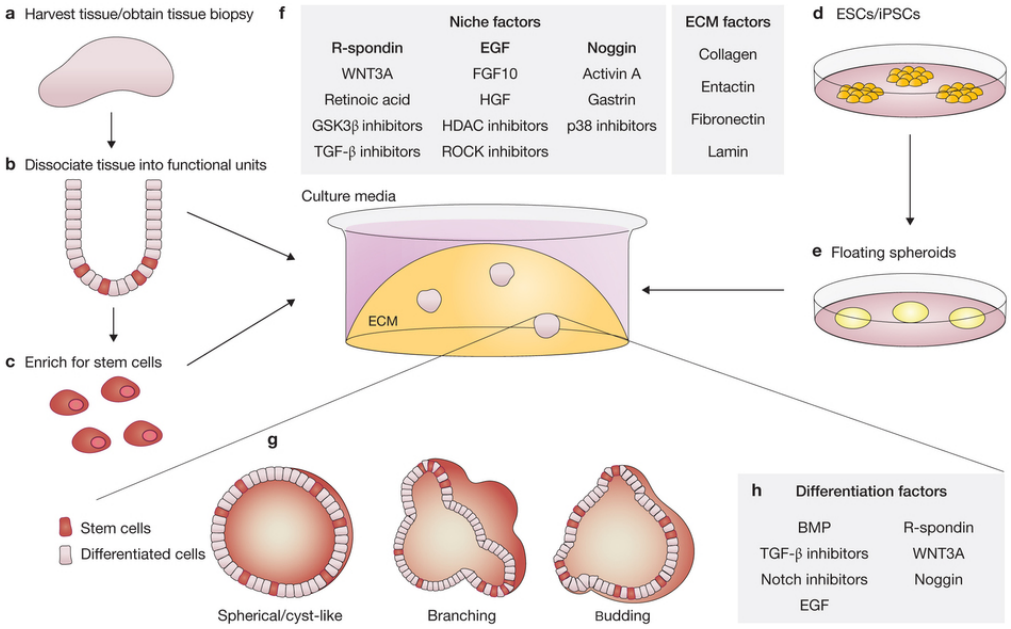

Figure 6.1: Overview of the general in-vitro organoid culture system. Adapted from Fatehullah et al. 2016

et al. 2015 noted that tumor-derived organoids presented with a range of patient-specific morphologies, from thin-walled cystic structures to compact organoids devoid of a lumen. It is therefore critical to develop a quantitative analysis that captures detailed organoid morphology consistent with current growth and imaging protocols.

6.2 Organoid Culture and Timelapse Imaging

To culture and grow organoids from primary tissue, the acquired tissue usually from biopsies in humans or dissections in mouse is dissociated into functional sub-tissue units containing stem cells and continually passaged to get single cells and enrichment of the stem cell fraction. Subsequently the mixture is embedded in extracellular matrix (ECM) surrounded by culture medium to initiate organoid culture. The culture is supplemented with organoid specific niche and ECM factors. Stem cells are maintained and perpetuated within the organoids, continually giving rise to differentiated progeny. Typical morphologies are classified as spherical, branching or budding. Organoids can either differentiate spontaneously or be induced to differentiate towards desired lineages or cell types by adding suitable

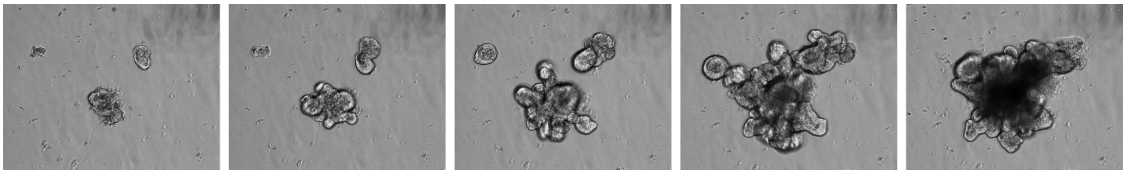

Figure 6.2: Monitoring of mouse intestinal organoid growth and development under timelapse phase-contrast microscopy for ∼1 week.

differentiation factors. Fig.6.1 illustrates this general process. Whilst successful for growth and maintenance of stem cells, the current culture protocols cannot grow organoids in a controlled fashion scalable to high-throughput multi-well plates. A single well will likely contain hundreds of organoids. Only select views, preferably containing single organoids will be taken for filming. Ideally the organoid could be fluorescently stained and imaged at single-cell resolution under timelapse with different coloured markers used to label the differentiated progeny. However this is extremely challenging at present. The presence of the ECM prevents efficient transfection, the depth the organoid is in the matrigel cannot be controlled, the organoids do not tolerate well the continual laser stimulation used in fluorescence imaging over long-times and cannot be fully sectioned optically through the entire volume. As a compromise, the grown organoids can be filmed using label-free phase-contrast or DIC (differential interference contrast) microscopy over long times (∼2 weeks, potentially unlimited depending on media changes) to monitor their growth and morphogenesis. However these acquisition methods do not offer high-quality optically sectioned z-stack images. The resultant z-stack acquisition are akin to an extended focus. Yet the organoid motion due to growth is three dimensional. In the rest of the thesis we refer to such image sequences of 3D phenomena acquired in or projected into 2D as '2.5D' acquisitions to distinguish them from pure '2D' sequences such as the two cell population videos of the previous chapter.

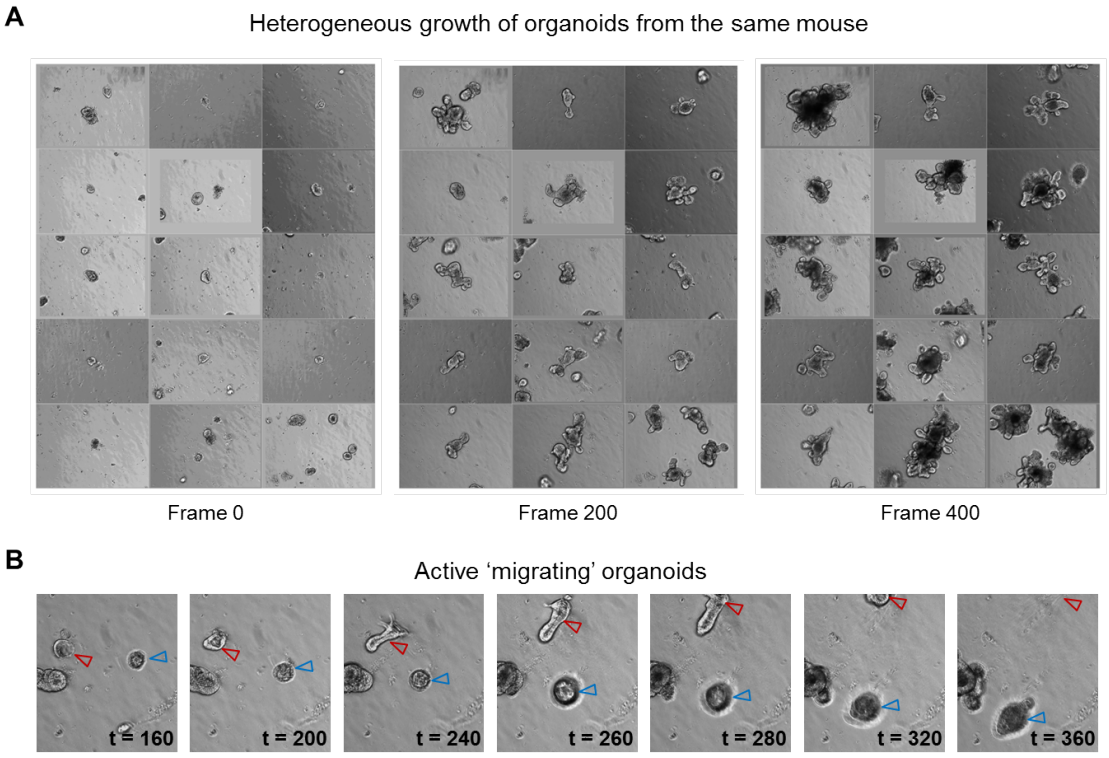

Figure 6.3: Variability of organoid culture. A) Heterogeneous development of intestinal organoids from the same wild type mouse. B) Example of actively migrating organoids in the matrigel indicated by red and blue arrows.

6.3 Technical Challenges for Automated Image Analysis of Organoids

The variability of current organoid cultures and timelapse imaging methods present technical challenges for the establishment of automated high-throughput image analysis platforms for organoid screening which can be grouped into specific imaging and analytical challenges. Fig.6.3 illustrates some of these challenges, where intestinal organoids were expanded from a single wild-type mouse with no genetic modifications.

Imaging Challenges

- Finite culture media requires changing over long-time filming. It is then difficult to relocate the same organoid and resume filming from the exact spatial location.

- Multiple organoids in the field of view
- Uneven and changing background illumination
- Out-of-plane, unfocussed organoids and debris
- Organoids can merge and overlap, occluding neighbouring organoids as they grow
- Migrating organoids and those not well embedded in the ECM, (Fig.6.3B).

Analytical Challenges

- Organoids filmed over different acquisitions need to be registered to ensure spatiotemporal continuity.
- Organoid depth is uncertain, size and morphology can drastically change, multiple optical z stack slices are required. The best slice need to be selected when compiling the video.
- Quantitative description and comparison of branching morphology.
- Quantitative phenotypic metrics that capture organoid heterogeneity.
- Detection and filtering out of necrotic, non-growing, migrating organoids and other 'outliers'.

In summary, large illumination variations between wells, large morphological changes and organoid occlusion combine to make it exceptionally challenging for frame-by-frame image segmentation based on image intensity and texture commonly provided by commercial and open-source image analysis software, (Fig.6.3). At the same time, the typically small number of videos in a screen (96 or 384 wells) and lack of training labels limits segmentation based on supervised deep learning approaches. Even suppose perfect frame segmentation of each organoid was possible, the resulting mask would not allow inference of cellular motion within the organoid shape in the resulting effective '2.5D' timelapse images. For these reasons, we

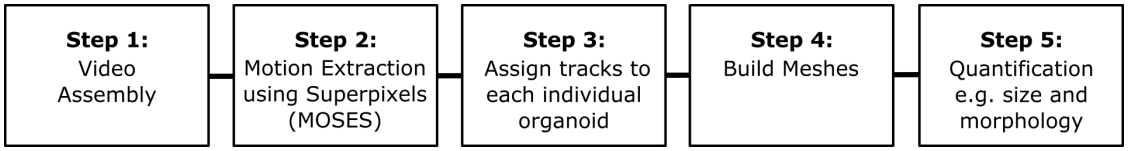

Figure 6.4: Overview of the automated organoid video analysis pipeline.

developed an automatic video analysis pipeline building upon that for migrating two-cell epithelial sheets in Ch.5.

6.4 Datasets

One timelapse video dataset was used to develop the automated pipeline and whose branching dynamics was subsequently analysed, (Ch.6.7). It comprises 80 videos of mouse intestinal organoids, acquired one frame every 15 mins over 7 days using a 10x objective. 4 conditions each with 20 videos are represented: 1) WT-wild type, 2)KO-p53$^{-/-}$, 3) WT+VC treatment, 4) KO+VC treatment. VC treatment comprises the addition of CHIR99021, a GSK inhibitor and valproic acid (VPA) and was used by Yin et al. 2014 to obtain homogeneous Lgr5+ intestinal stem cells (ISC). Together VPA and CHIR99021 together denoted VC maintains ISCs in an undifferentiated, self-renewing state and leads to distinctive drastic changes in the external organoid morphology.

6.5 Automated Video Analysis of Organoid Development using MOSES

6.5.1 Overview of Pipeline

The overview of the developed automated analysis pipeline for organoids is illustrated in Fig.6.4. Organoid analysis using MOSES is similar to that of the analysis of two migrating epithelial sheets (Ch.5) but presents new and additional challenges that requires the development of new algorithms. First, the need to acquire timelapse videos over several acquisitions due to growth medium renewal necessitates the usage of automatic image registration to realign frames for spatiotemporal continuity (Fig.6.4 Step 1). Second, the presence of multiple organoids introduces multiple

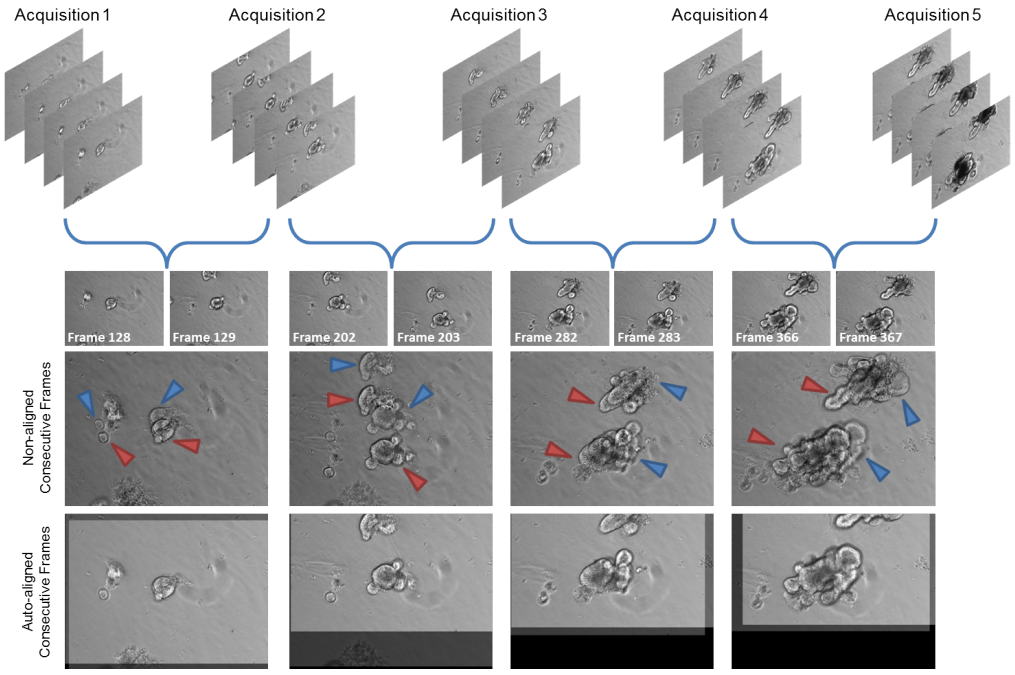

Figure 6.5: Automatic alignment of organoid video frames from multiple imaging acquisitions.

'objects' requiring a method to assign superpixel tracks specific to each organoid (Fig.6.4 Step 3). This is further complicated by organoids that grow into each other. Previously the epithelial sheet was only one 'object'. Third, estimation of overall organoid shape frame-to-frame and finally, the detection and tracking of branching dynamics from the superpixel tracks (Fig.6.4 Step 5). Below we describe solutions to each of these components.

6.5.2 Automated Image Registration of Video Frames

Organoid development from primary tissue is slow. To reliably capture the full development of branching such as that in intestinal organoids, long-time acquisition over one to two weeks is required. As the organoid develops over this time frame, the surrounding growth medium is depleted and must be replenished. The entire timelapse sequence must therefore be acquired over separate acquisitions, (Fig.6.5). Relocation of the same organoid and resumed filming from the same spatial position is often imprecise as shown in Fig.6.5 after automatic selection of the best focussed z slice per frame using the variance of laplacian (LAPV, Pertuz et al. 2013)

with blue arrows indicating the last frame of the previous acquisition and red arrows the first frame of the subsequent acquisition. Established Fourier Transform correlation methods (Guizar-Sicairos et al. 2008) is applied sequentially to correct the translational shift between successive image frames with the aligned frame serving as the template for the next frame , (Fig.6.5 bottom-most panel). A final global alignment was not needed. This realignment is vital to avoid the introduction of artificial motion prior to the application of MOSES for motion extraction.

6.5.3 Instance Segmentation of Organoids

Ideally, the imaging field of view contains only a single organoid. However this is not always possible in practice due to stochastic variation in the resulting spatial placement following dissociation and matrix embedding with current organoid culture protocols. In addition label-free phase contrast microscopy in contrast to fluorescence microscopy captures both the desired organoids and a non-relevant image background that comprises for example the movement of cell debris and imperfections in the matrix. Naive 'out-of-the-box' motion extraction using optical flow sufficient for monitoring the migration of epithelial sheets here result in noisy motion tracks that are suboptimal for further analysis, (Fig.6.6A). To overcome this limitation each initial organoid must be uniquely identified and for each, specific motion tracks must be assigned. This is known as 'instance segmentation' ("where are the individual organoids") as opposed to 'classification' ("Are there organoids in the image") and is a very challenging problem in general (c.f. Ch.2.3.3). Here we solve this problem by detecting the (x, y) position of each organoid in the first frame and retaining those organoids which contribute to the global motion by attempting to match the detected organoid centres with the inferred motion centres of superpixel tracks. Superpixels are subsequently assigned based on analysing the spatial connectivity of moving superpixels. The core steps are illustrated in Fig.6.6.

Step 1: Forward and Backward Motion Extraction. The first frame at time, t_0 is partitioned into a regular number of superpixel regions as described (Fig.4.1) and tracked frame-by-frame forward in time from t_0 to the final

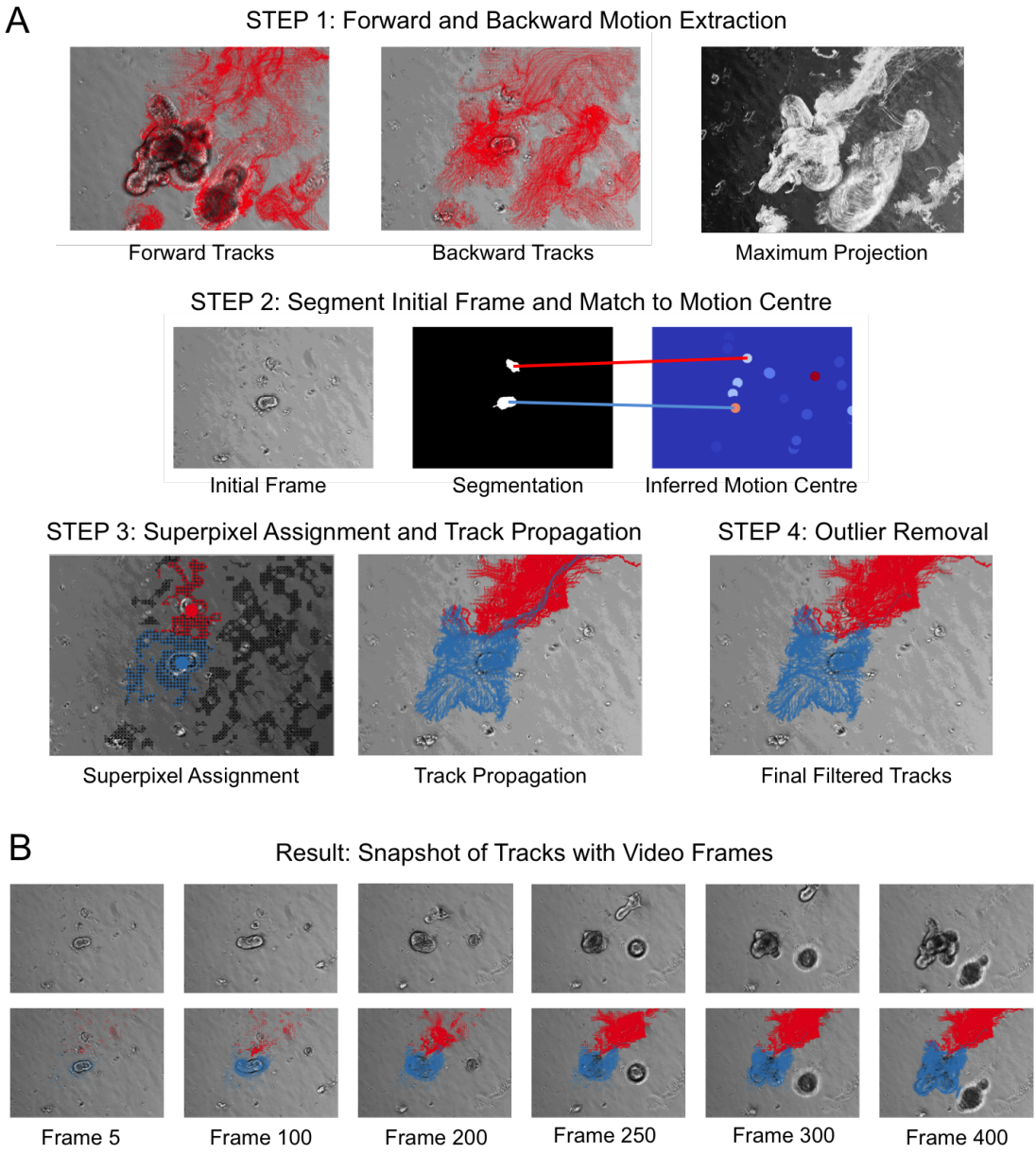

Figure 6.6: Automated pipeline for single organoid motion extraction in the presence of multiple organoids. A) Overview of the proposed pipeline (steps 1-4). B) Overlay of final identified motion tracks onto video frames up to the specified video frame number i.e. frame 100 is the position of each track up to frame 100. Different colour indicates a different organoid.

frame t_T where T is the total number of video frames. The set of (x_t, y_t) superpixel positions form the forward superpixel tracks. Similarly the final video frame at time t_T is partitioned into a regular number of superpixel regions and tracked frame by frame backwards from t_T to t_0 to extract the set of (x_t, y_t) superpixel positions to form the backward superpixel tracks. The forward tracks reveal where motion moves towards or 'motion sinks', (Fig.6.6A Step 1 left), whereas the backward tracks reveals where motion originates or 'motion sources', (Fig.6.6A Step 1 middle), in analogy to fluid flow (Ch.3.2.2). Combined they describe the full range of dynamic motion in the movie as can be visualised using the maximum projected image of the entire video, (Fig.6.6A step 1 right).

Step 2: Segment Initial Frame and Match to Motion Sources. The initial frame is segmented to identify each organoid of a minimum size (specified by the number of pixels), (Fig.6.6 step 2 middle). This segmentation is at present dataset specific. Currently an intensity-based segmentation is employed. The image background is first estimated using morphological closing and subtracted. Non-local means (Buades et al. 2005) is then applied for image denoising. To segment, GMM clustering (2 components) is applied and the resulting binary mask is binary infilled and connected components smaller than a minimal size are removed to obtain the final foreground organoid mask (Fig.6.6A, step 2). To infer a likelihood image of possible motion sources, the MOSES saliency map using the most sensitive radius neighbour mesh ($r \leq 5w_s$) (c.f. Ch.4.3.2, 4.4.3), where w_s is the average superpixel width is computed. A local maximum filter (disk kernel, radius $2w_s$) is applied to the result followed by non-maximum suppression (setting all pixels not a local maximum within a local circular neighbourhood of w_s pixels to 0) to minimise false detections (Fig.6.6 step 2 right). Organoids are assigned to detected motion sources if their respective centroids differ no more than a distance given by a specified constant multiple (1.5) of the equivalent circular radius computed from their binary mask, (Fig.6.5.3 step 2). Each organoid

is associated with a single motion source but each motion source can be associated with multiple organoids. Organoids not matched to any motion source is regarded as non-moving for the entire video and removed from further consideration.

Step 3: Superpixel Assignment and Track Propagation. Determine the set of initially moving superpixels, $\mathcal{S}_{move}^{t_0}$ defined as those superpixels whose summed displaced distance between successive frames from the initial frame $t = t_0$ to the frame of movement t_m is greater than 0,

$$\left\{ \sum_{t=t_0+1}^{t_m} |\mathbf{r}(x_t, y_t) - \mathbf{r}(x_{t-1}, y_{t-1})| \right\} > 0$$

t_m is automatically determined for each video and is defined as the smallest frame number $t_m = \min(1, t_{\text{move}})$, that satisfies $t_{\text{move}} = \arg\min_t F[t] = \left(\sum_{i=1}^{t} d[i]\right) > d_c$, for $i = 1, 2, \ldots, t$ where $d[i] = \frac{1}{N}\sum_{j=1}^{N} |\mathbf{r}_j(x_i, y_i) - \mathbf{r}_j(x_{i-1}, y_{i-1})|$, $\mathbf{r}_j(x_i, y_i)$ is the (x, y) position of superpixel j at time frame i and d_c is a user-set cut-off distance (1.5 pixels). $F[t]$ is essentially the cumulative mean superpixel displacement between frames from the initial frame 0 to frame t. Then form the region adjacency graph (RAG) between superpixels in $\mathcal{S}_{move}^{t_0}$ using the initial (x, y) positions at $t = t_0$ and find connected components of the RAG. Connected components of a graph are defined as a subgraph whose any two vertices or nodes are connected to each other by an edge but is connected to no other vertices outside of the subgraph belonging to the larger graph. Thus a graph whose nodes are fully connected has only one connected component. Each retained organoid centroid is assigned to a graph component based on euclidean distance. Fig.6.5.3 step 3 left, illustrates this visually. The red and green centroids form a single graph component in which two organoid centroids have been assigned to whilst the black centroids are graph components in which no organoid is assigned. If more than one organoid centroid is assigned to a single graph component, then the respective graph component is 'split' such that each superpixel centroid comprising the graph component is uniquely associated to the nearest organoid centroid as

measured by the euclidean distance, (Fig.6.5.3 step 3 left). Effectively this assumes that each organoid (and its motion range) can approximately be modelled by a Gaussian prior. To handle deviations from this assumption (e.g. due to occlusion between organoids and organoid branching), this initial assignment is semantically refined through pairwise 2-component GMM (Gaussian mixture model) clustering on the PCA dimensionally reduced superpixel tracks (Ch.4.4.1) . Each organoid i, O_i now has associated a unique set of superpixels $\mathcal{S}^{O_i}_{\text{initial}} \subseteq \mathcal{S}^{t_0}_{\text{move}} \subset \mathcal{S}$ whose tracks in time capture the dynamic motion of organoid i. However, only the initially superpixels that were moving were considered in this assignment. There is a high probability of superpixels that initially do not move but become 'activated' as the organoids grow in size similar to the migration of the epithelial sheets in Ch.5. To improve the accuracy of the full dynamic capture for each organoid O_i, we additionally augment O_i with the subset $\mathcal{S}^{O_i}_{\text{activate}}$ which are superpixels that are a member of $\mathcal{S}^{t_0}_{\text{no move}} = \mathcal{S} \setminus \mathcal{S}^{t_0}_{\text{move}}$ but become 'activated'; where the initial (x_0, y_0) position of $S_j(x_t, y_t) \in \mathcal{S}^{O_i}_{\text{activate}}$, $S_j(x_0, y_0)$ comes to lie within the combined total spatial area covered by the set of initial superpixels $\mathcal{S}^{O_i}_{\text{initial}}$ over all time, $\Omega^t := \{\omega^t_1, \omega^t_2, ...\omega^t_k\}$ where ω^t_k is the area of superpixel k at time t. This is similar to the propagation when analysing the motion of two cell populations (Ch.5.3.1) except the 'activation' is not applied sequentially for each frame. The final set of superpixels and associated tracks that capture the dynamic motion of organoid O_i is the union of the initial and activated superpixels, $\mathcal{S}^{O_i} = \mathcal{S}^{O_i}_{\text{initial}} \cup \mathcal{S}^{O_i}_{\text{activate}}$ (Fig.6.6 step 3 right).

Step 4: Outlier Removal. (Optional) When organoids are initially close spatially or become close during filming through growth or movement there is a high probability of some misassigned tracks, (Fig.6.6 step 3 right) though the majority is correct. The presence of these 'outlier' tracks could affect downstream analysis such as branch detection (Ch.6.6.3) and should be filtered out. This can be efficiently achieved by exploiting the idea of point density to find the total spatial range of motion (c.f Ch.4.6.1) and to use this cue

to remove outlier superpixel tracks whose motion is not constrained within this spatial range. For organoid O_i described by n_s superpixel tracks $\mathcal{S}^{O_i}$, collapse the time t and superpixel identity k to form the super point set, $\mathcal{P}^{O_i} := \{(x_{i'}, y_{i'}) | (x_t, y_t, k) \mapsto (x_{i'}, y_{i'}), i' = 1, .., N\}$ where $N = K \times T$ is the total number of points, for K superpixels and T video frames. Then map the full point set $\mathcal{P}$ into a spatial image, I_P whose pixel values at (m, n) where m, n are the row and column indices is the number of points at that location:

$$I_P(m,n) = \begin{cases} \sum_{i'}^{N} \mathbf{1}_{x_{i'}, y_{i'}} & , \text{if } y_{i'} = m, x_{i'} = n \\ 0 & , \text{Otherwise} \end{cases} \tag{6.1}$$

The resulting image is sparse relative to the number of image pixels, depending on the initial number of superpixels specified. To 'densify' the image for segmentation, the image I_p is smoothed with a Gaussian kernel ($\sigma = 2w_s$) and binarised by 3 cluster K-means (retaining the highest intensity 2 clusters) to produce a binary mask, whose largest connected component is then retained to capture the dynamic motion range as it is assumed superpixel tracks belonging to any single organoid should move as a single entity and be 'close' to its neighbours. Each superpixel track k, $S_k(x_t, y_t) \in \mathcal{S}^{O_i}$ is then assigned a reliability score, R^k_{score} defined as the fraction of time its motion is within the spatial region covered by the final binary mask, BM_{motion} where $R^k_{score} := \frac{1}{T} \sum_{t=t_0}^{T} BM_{\text{motion}}(y^k_t, x^k_t)$. A second reliability measure is the end point error, EPE^k defined as the minimum distance of the final position of the superpixel k with respect to the pixels in BM_{motion}. A superpixel track k is categorised as reliable and kept if it satisfies the following dual criteria:

$$R^k_{score} > R_{\text{thresh}} \quad \wedge \quad \text{EPE}^k < \mathbb{E}_k[\text{EPE}^k] - 2\,\text{Std}_k[\text{EPE}^k] \tag{6.2}$$

where $\mathbb{E}, \text{Std}$ is the mean and standard deviation over all superpixels k, $R_{\text{thresh}} = 0.5$ is a user specified threshold and $\wedge$ the logical and operator. The final set of superpixels removes outlier tracks whilst minimally affecting inlier tracks, (Fig.6.6 step 4). A more natural choice of statistic for robustness is the median but we found the median to estimate too low the upper bound and

eliminates too many tracks. In practice this outlier removal step is quite time consuming and was only necessary in a handful of videos where two organoids develop very closely to each other and both exhibit significant deviation from a 3D isotropic shape due to branching as might be the case for intestinal organoids. As such it is deemed an optional post-processing step.

The final superpixel tracks isolate the motion of individual organoids and partitions the movement spatiotemporally. Plotting the organoid superpixel tracks in different colours from the initial time $t = t_0$, frame 0 to frame 400, the superpixel tracks record the time history of the motion, providing easy visualization of the number of organoids and their movement in one static image which could not have been inferred from a single or a few frames, (Fig.6.6B).

6.5.4 Tracking of Organoid Morphology

In cases where the predominant motion of the organoid is growth, the superpixel tracks allows for inference of the individual organoid morphology from the '2.5D' timelapse sequences. Both identification of the external shape as well as tracking of the individual branching dynamics as the organoid develops can be recovered to allow further quantification.

Estimation of Organoid Shape

Precise segmentation of the external organoid shape frame-by-frame from sparse superpixel tracks is difficult where the organoid is represented by a much larger number of pixels than the fixed number of superpixels used for motion extraction. Nevertheless, the spatial distribution of the superpixels from forward and backward tracking can provide an estimation of the organoid shape comparable to pixel-based image segmentation, (Fig.6.7A). Similar segmentation from trajectories has been applied to segment general objects in videos, (Brox and Malik 2010; Fragkiadaki et al. 2012). Our basic workflow is as follows:

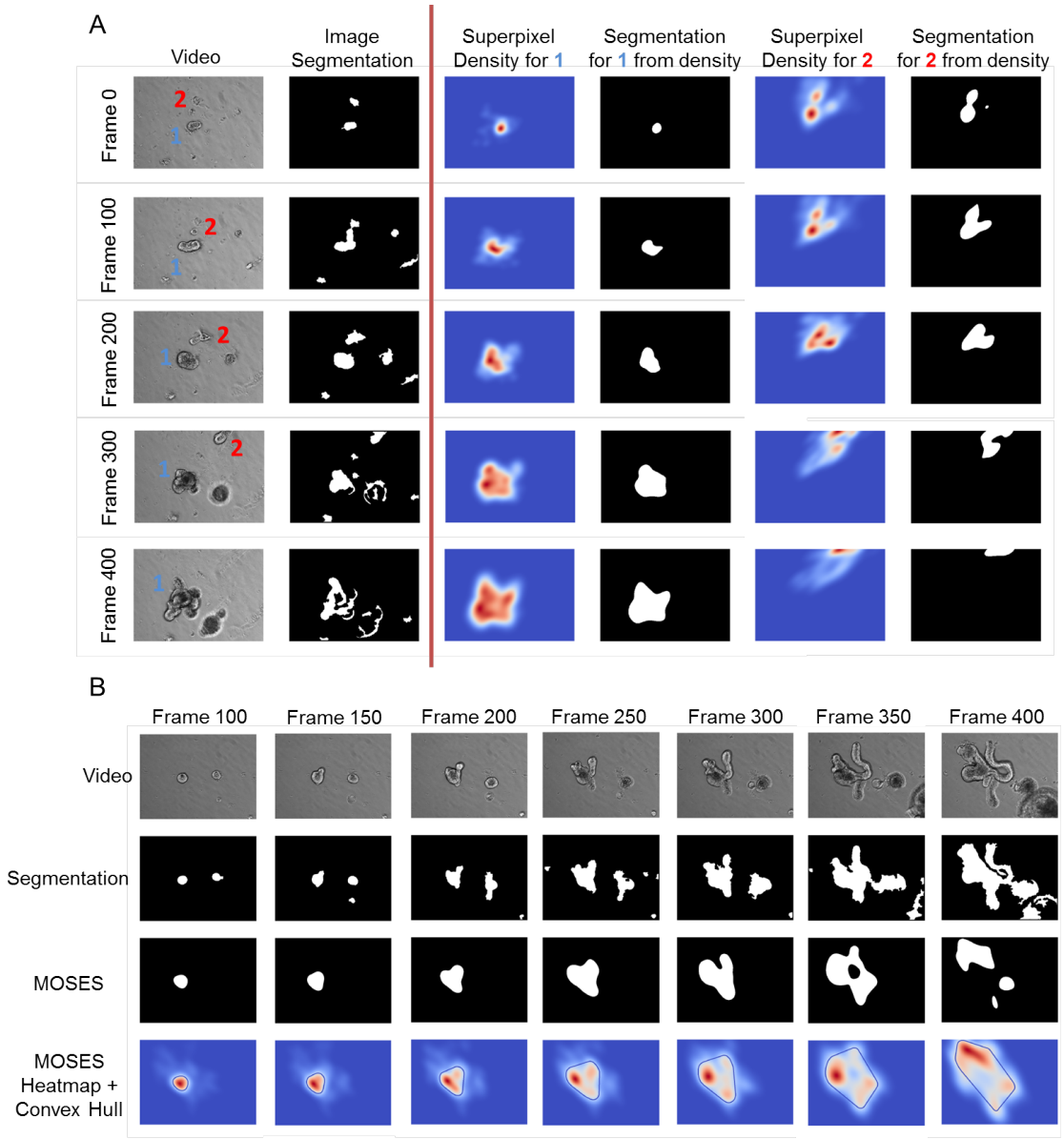

Figure 6.7: Estimation of organoid shape from motion tracks. A) Heatmap and corresponding binary segmentation for each organoid in comparison to image segmentation for frames 0, 100, 200, 300 and 400 for organoid 1(blue) and organoid 2(red). B) Snapshots of a second video comparing shape segmentation from motion using MOSES with image pixel intensity based segmentation (2-component GMM clustering on the denoised background subtracted image (see Ch.6.5.3, step 2)).

- Given a time frame t ($t > 0$), concatenate the forward and backward superpixel positions $S^F(x_t, y_t)$, $S^B(x_{-t}, y_{-t})$ to form the combined point set $P^{FB}(x_t, y_t) = S^F(x_t, y_t) \cup S^B(x_{-t}, y_{-t})$.

- Gaussian kernel density estimation with automatic bandwidth selection using Scott's Rule (Scott 2015) is applied to estimate the 2D probability distribution of the set $\mathcal{P}^{FB}(x_t, y_t)$, generating a dense spatial heatmap for visualization, (Fig.6.7).

- The heatmap is binarised using 3-class K-means with the highest intensity class retained to estimate the organoid shape with a binary mask. Two classes, (foreground/background) was found to overestimate the organoid shape in early frames.

This approach works well for small organoids or over the majority of the video (total 468 frames) for a large organoid, (Fig.6.7A,B) when we compare to the difficulties of frame-by-frame image segmentation. However for large organoids, in the latter frames over a long-time acquisition, growth slows and there may be greater cell death. The entire organoid will no longer undergo motion changes in consecutive frames. Here the above binarisation results in undersegmentation of the organoid, primarily identifying only the growing branches, (Fig.6.7B, MOSES panel frame 400). Nevertheless an extra step, computation of the convex hull image of the binarised heatmap, shown by the outline around the heatmap in Fig.6.7B, bottom panel, ensures robust estimation. In the worst case it allows recovery of the bounding box. The convex hull of a binary image is defined as the set of pixels included in the smallest convex polygon that surround all white pixels in the input.

For primarily migrating organoids like organoid 2 in Fig.6.7A, the developed method accurately reflect the non-growing nature; maintaining a similar number of white pixels with gradually reducing area as the organoid disappears out of the field of view. The residual red in the heatmap is because individual superpixels are

not allowed to leave the field of view by design to maintain a constant number of superpixels. Thus the (x, y) positions are bounded by the size of the image.

Detecting Branching Processes

The set of organoid superpixel tracks, $\mathcal{S}_i^O$ not only captures the overall shape of organoid O_i but additionally captures information about the developmental dynamics of branching. Visually it can be seen that manual annotation of organoid branches is positively correlated to the number of tracks passing through the same spatial region of interest, (Fig.6.8A). We refer to tracks in this spatial zone as 'densest' tracks. Conceptually, branching can be viewed as a locally accelerated expansion of a group of budding cells compared to its neighbours and this appears to be captured in the superpixel motion tracks. In organoids that experience pure growth such as organoid 1, the 'densest' tracks can recover the branching dynamics (Fig.6.8A). In organoids of pure movement such as organoid 2, the 'densest' tracks provide a compact summary of the motion pattern. For fair comparison of branching dynamics, only purely growing organoids are retained for branching analysis. They can be automatically detected by filtering based on the angular spread of motion (growing organoids grow in all directions) and by shifts in the organoid centroid.

Based on the empirical observation relating organoid branching with the spatial concentration of motion tracks, an automatic branch extraction algorithm was developed that proposes branch points are spatial enrichment regions in the dynamic mesh and are local maxima ('blob-like hotspots') that can subsequently be tracked and joined together across frames using particle tracking or single cell tracking approaches to form branch tracks c.f. Fig.6.8B.

Step 1: Detect Local Maxima and Link Frame to Frame. For frames 1 to T (the final video frame), a heatmap localising the spatial region with the densest number of superpixels is given by the radius neighbour mesh MOSES saliency map ($r \leq 1w_s$) (Ch.4.3.2, 4.4.3) constructed from the forward superpixel tracks where w_s is the average superpixel width. Peaks given

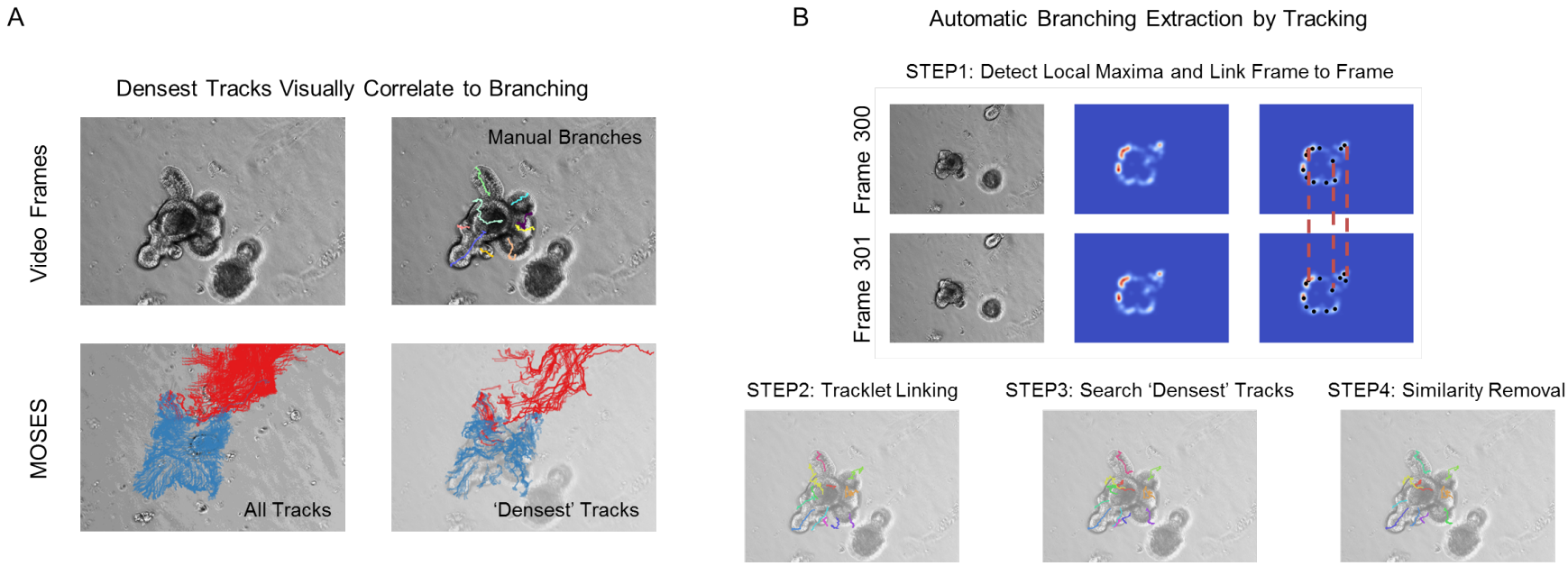

Figure 6.8: Tracking organoid branch development through motion using MOSES. A) Snapshot of the last frame, frame 468 showing final organoid branch shape with and without overlay of manually tracked branches. Each unique branch is a different colour, top panel. All MOSES motion tracks for each organoid and 'densest' spatially enriched motion tracks for visual comparison to the corresponding images above, bottom panel. B) Overview of the automatic branch extraction pipeline, different colour for each unique branch. Colours are not matched to the manual branches.

by (x, y) coordinates are identified using a local maximum filter and linked frame to frame using a simple nearest neighbour approach with a distance cut-off of $2w_s$. If a track cannot be linked to a point in the next frame, it is terminated. Points that could not be linked to a point in the previous frame are initiated as new tracks.

Step 2: Tracklet Linking. The frame-to-frame linking process produces a series of tracks of varying time lengths termed tracklets. After discarding very short tracklets, tracklets are grouped according to their start time and groups are sorted by time in ascending order. Starting from the first group, tracklets are tested successively for linking based on distance i.e. if the last location of tracklet A in group 1 is within a distance $d_{\text{link}} = 3w_s$ of the first location of tracklet B in group 2 then they are linked. To link tracks, a new track is produced with the start time of tracklet A and the end time of tracklet B. Where tracklet A and tracklet B overlaps in time, the mean of the two positions is computed. Tracklets that cannot be linked with any track in the next group is deemed to be unique and not considered for further linking. This linking process is necessary to combine tracklets that may belong to a longer track but was not identified in every frame due to image artifacts such as occlusion disrupting the motion signal.

Step 3: Search the Set of 'Densest Tracks'. The resultant tracks, $\mathcal{T}_{\text{link}}$ based on local maxima detection and tracklet linking may still result in shorter or longer tracks than expected depending on the quality of the linking which can vary from video to video, (Fig.6.8B step 2). Further it may not agree with the motion captured in the superpixel tracks. Instead we view the $\mathcal{T}_{\text{link}}$ as 'queries' to retrieve the most representative track from the set of 'densest' tracks, $\mathcal{D}_{tracks}$ (superpixel tracks that have many neighbouring tracks as depicted in Fig.6.8A) which we view as a pool of 'templates'. The measure for retrieving the most similar track for track $\mathcal{T}_i$ is to select the track $\mathcal{T}_j \in \mathcal{D}_{tracks}$ that minimises the mean euclidean distance, $d(\mathcal{T}_i, \mathcal{T}_j)$. Formally, the following

density score function, $D_{\mathcal{T}_i}$ is used to score the 'density' of a superpixel track $\mathcal{T}_i$ and measures the average fraction of time neighbouring tracks of $\mathcal{T}_i$ lies within a constant distance cut-off, d_c set using either prior knowledge or through statistical means by regarding each superpixel as a 2D Gaussian (as described in Ch.4.3.1).

$$D_{\mathcal{T}_i} := \frac{1}{|\mathcal{N}_i|} \sum_{\mathcal{T}_j \in \mathcal{N}_i, j \neq i} \mathbb{E}[d(\mathcal{T}_i, \mathcal{T}_j) < d_c] \tag{6.3}$$

where $\mathcal{N}_i$, is the set of neighbour superpixel tracks of $\mathcal{T}_i$ defined as all superpixels within a distance of d_c in the first frame t_0, $\mathcal{N}_i := \{\mathcal{T}_j | d(\mathcal{T}_i(x_0, y_0), \mathcal{T}_j(x_0, y_0)) < d_c, j \neq i\}$. $\mathbb{E}$ is the mean taken over the trajectory length. Otsu thresholding applied to the set of all $\{D_{\mathcal{T}_i}\}$ produces a constant cut-off D_{thresh}. All superpixel tracks whose density score $D_{\mathcal{T}_i} > D_{\text{thresh}}$ is considered dense, i.e.

$$\mathcal{T}_i \in \mathcal{D}_{tracks} \quad \text{if} \quad D_{\mathcal{T}_i} > D_{\text{thresh}} \tag{6.4}$$

This sequence of computing the score function and assignment is independently applied 3 times to the forward and backward tracks iteratively to refine the density candidates with $d_c = w_s$ (the average superpixel width).

Step 4: Removal of 'Similar' Tracks. The resultant tracks may contain tracks with similar shape that overlap spatially or tracks whose end positions colocalise. These tracks with high probability belong to the same organoid branch and should be described by a single representative track. Instead of taking the average of these tracks, the track who has moved furthest in the final frame at time T, (x_T, y_T) from the initial position, (x_0, y_0) is designated as the representative track. Track similarity was defined as before with the trajectory cross-correlation (TCC) (with values between 0 and 1), (Ch.4.5, Eqn.4.18). Spatial overlap of a track $\mathcal{T}_i$ and track $\mathcal{T}_j$ is defined by the maximum overlap ratio of the convex hull image of the respective tracks denoted ConvHull($\mathcal{T}_i$) and ConvHull($\mathcal{T}_j$).

$$\text{SO}[\mathcal{T}_i, \mathcal{T}_j] = \max \left\{ \frac{\text{ConvHull}(\mathcal{T}_i) \cap \text{ConvHull}(\mathcal{T}_j)}{\text{ConvHull}(\mathcal{T}_i)}, \frac{\text{ConvHull}(\mathcal{T}_i) \cap \text{ConvHull}(\mathcal{T}_j)}{\text{ConvHull}(\mathcal{T}_j)} \right\} \tag{6.5}$$

If two tracks $\mathcal{T}_i$, $\mathcal{T}_j$ have a TCC$\geq$ 0.95 spatial overlap, SO $\geq$ 0.7, they are marked for combining and the representative track is found as described above. For a cluster of tracks (>2) that are similar (TCC$\geq$0.95), all pairwise spatial overlaps are computed. The tracks with the largest convex hull images are considered and the representative track is computed from this set as described above.

6.6 Validation of Pipeline

There is no suitable benchmark dataset available for assessing the quality of motion extraction for organoids. Public datasets for assessing optical flow such as Middlebury, KITTI and Flying Chairs (Ch.2.1.3) are not suitable. They exhibit larger displacements than biological motion and is of everyday objects which possess very different image statistics. There is also no validated method to generate synthetic data. Finally, it is too laborious to outline frame-by-frame organoid shapes for a large number of organoids. Each intestinal organoid video for example is 468 frames, often with many organoids in one image, (Fig.6.9A). Instead we verified the pipeline indirectly using three tests that assess the quality of the results inferred from the organoid superpixel tracks, for which less laborious annotation could be completed by multiple human annotators.

- Test 1: Manual visual assessment of organoid tracks
- Test 2: Validation by counting
- Test 3: Validation of organoid branching dynamics

Tests 1 and 2 verify the ability to accurately identify moving organoids and correctly partition its spatial area of motion. Test 3 indirectly verifies the accuracy of the motion extraction. Only if the motion is extracted faithfully can the inferred branching tracks correlate with manually identified branches. Further it should perform significantly better than randomly choosing from the pool of dense tracks which we use as a baseline algorithm for comparison.

6.6.1 Manual Visual Assessment of Organoid Tracks

Due to the lack of ground-truth data and challenges in generating an applicable synthetic dataset with suitable ground-truth for organoids, the quality of the motion assignment was visually assessed. Examples of track assignment is shown for 8 videos in Fig.6.9A demonstrating its performance in the presence of single and multiple organoids. Tracks were overlaid on static video snapshots and compared with sequentially sampled frames. Combined with Fig.6.7, one of the major evident benefits of the MOSES motion tracking approach is the ability to leverage information in previous frames to ensure spatiotemporal continuity of organoid identification and separation of individual organoids even in highly confluent environments where they may be physically touching or occluding each other in the latter half of the video. Such frames would otherwise be extremely difficult if not impossible to identify with frame-by-frame image segmentation.

6.6.2 Validation of Segmentation Through Counting

Three annotators separately counted the number of initial organoids that underwent 'significant' movement over the course of the video. Under this criteria, new organoids moving into the field of view after the initial frame are not counted. Those organoids that only move a small distance before undergoing apoptosis are also not counted but initial out-of-focus organoids that grow and thereby exhibit significant movement are counted. Applying this criteria, 77/80 videos contained moving, live organoids. The automatic number of distinct organoids was compared with the average number of organoids of the 3 annotators. The counts exhibited strong linear correlation with a Pearson correlation coefficient of 0.867 and a slope of 0.895 where the ideal is 1 indicated by the dashed black line. Assessing the per video difference to human annotations, we report 100% agreement in 73% (56/77) videos. For comparison, humans achieve 100% agreement in 92% (71/77) of the videos. In reality for screens one would aim to establish a lower density of organoids, ideally one organoid per field of view. For 21 of the videos where

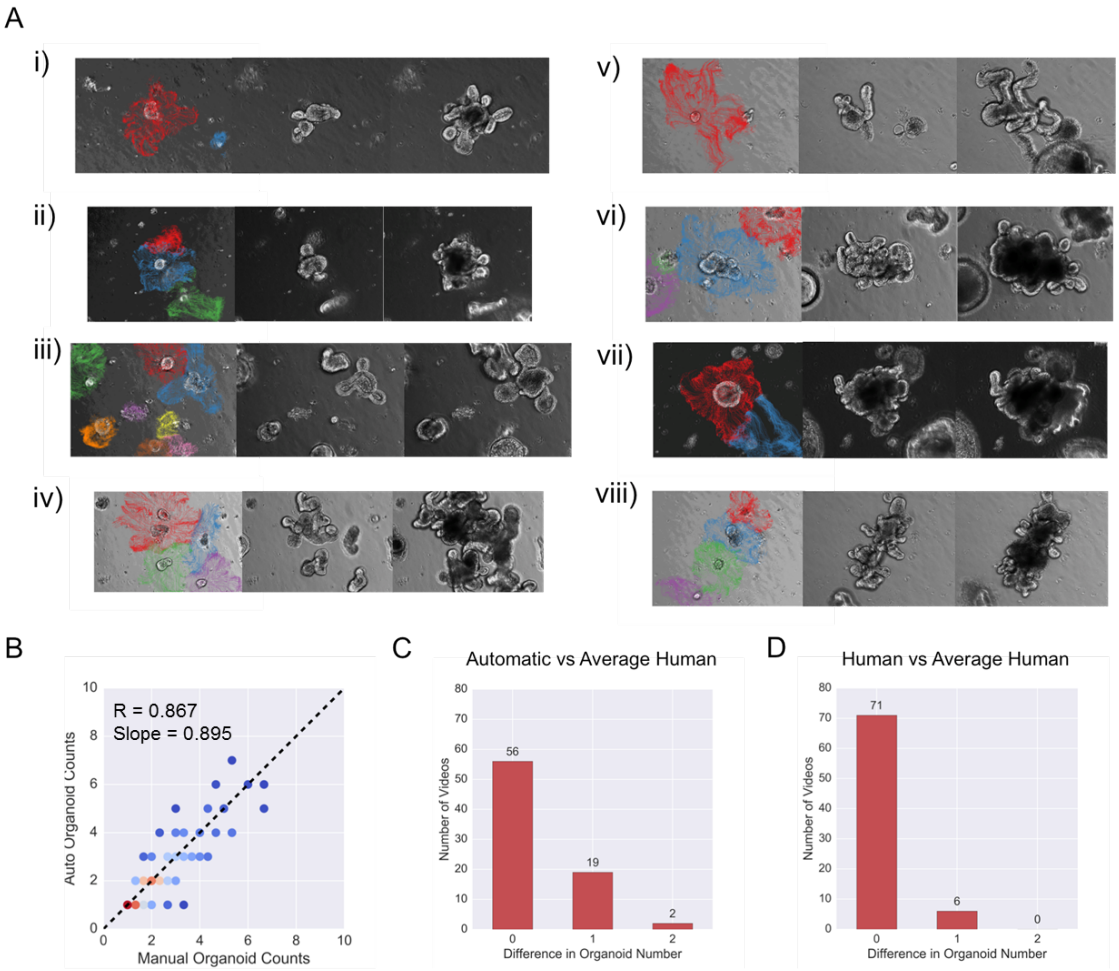

Figure 6.9: Qualitatitive and quantitative assessment of organoid segmentation from motion. A) Snapshots of motion track assignment for organoid in 8 different videos i)-viii). For each sequence of three images, the first displays the assigned motion tracks overlaid on the initial video frame, the second a snapshot from the middle of the video to show movement, the third is the final video frame of the video. B) Plot of the number of automatically identified moving organoids vs the average number of manually annotated moving organoids from 3 human annotators. C) Histogram of the difference in organoid number rounded to the nearest integer between the automated count and the average human count. D) Histogram of the average difference in organoid number rounded to the nearest integer between each of the three human annotators and the average human count of the three annotators.

only a single organoid was moving (agreed on by all 3 annotators), the proposed algorithm achieves 100% agreement with humans.

6.6.3 Validation of Branching Through Detection

It is impractical and ambiguous to draw out the shapes of each organoid branch manually per frame for each video. Instead we compared the automated tracking to manual annotation of the branch tips frame-by-frame for a subset of 9/80 videos sampled uniformly from each of the 4 conditions. The annotated organoids were chosen on a traffic light system of 'easy' to 'challenging' based on qualitative assessment of the degree of movement of the organoid body, number of sprouting branches and linearity of the branching development, (Fig.6.10A). The automatically produced tracks was then compared to the manually annotated tracks in three ways i) the agreement in branch counts, ii) the similarity in the track shape after matching to annotated tracks based on initial starting position and iii) the temporal disagreement in the branch growth duration measured by the average difference in video frames. For ii) and iii) we also compare the performance of the branch tracking against randomly assigning the matched number of tracks from the set of dense tracks and the respective assigned organoid superpixel tracks. The KO-VC-1 video in Fig.6.10A was excluded for quantitative validation as both organoids were annotated in the same video and it exhibited extreme deviation from normal organoid development. We find strong positive correlation (Pearson correlation coefficient, 0.93) with the manually determined number of tracks irrespective of difficulty, with a small systematic overestimation (slope of 1.18 compared to the ideal of 1). On average the difference in branch counts per video was 4.0±1.1 branches with an average of 13.5±2.3 branches, a mean accuracy of 70.8% (Fig.6.10B,C). This overestimation is due to the use of a fixed number of superpixels for motion extraction. As the organoid grows, the fixed number of superpixels become dispersed over a greater spatial area and are concentrated at the boundaries. Consequently the density of tracking is reduced particularly with respect to tracking the dynamics within the growing organoid body. Without sufficient number of neighbours the

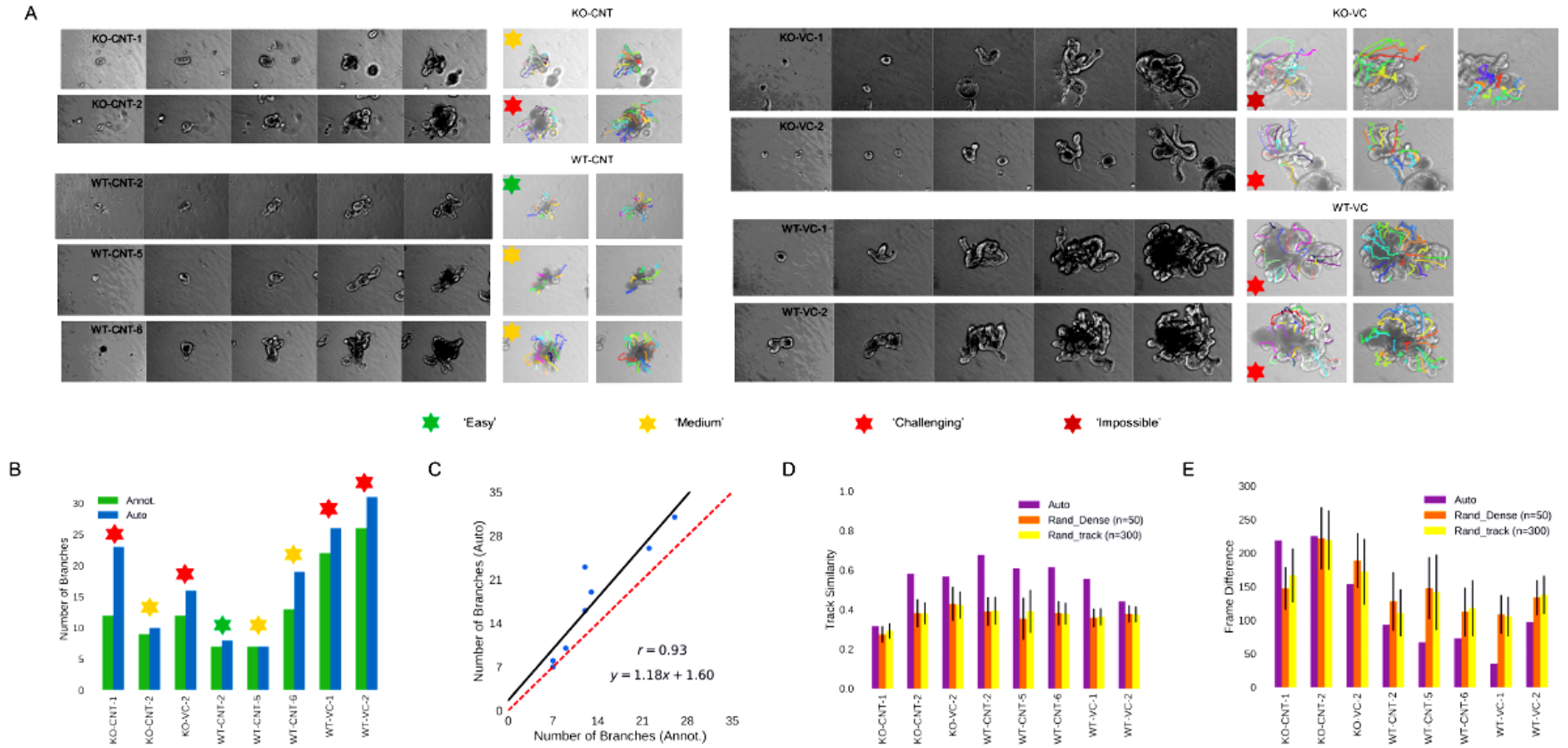

Figure 6.10: Qualitative and quantitative assessment of organoid branch development from motion. A) Static video snapshots of each of the manually annotated 9 videos at frames 0, 100, 200, 300, 400. Manually labelled branch tracks are on the left, automatically derived tracks on the right. Tracks are overlaid on the final video frame. Unique branches are labelled a different colour. The video KO-VC-1 contains two organoids that are initially separate. Thus MOSES produces two independent branch tracking results but the manual annotation was done for both on the same video due to the difficulty of annotating the organoid separately in later frames. MOSES can produce individual branch tracking for each organoid results are reported by MOSES. Each video was assigned a difficulty based on a traffic light star system. B) Bar plot showing the counted number of branches for manually annotated (Annot.) and automatically generated (Auto.) with MOSES. C) Plot of organoid numbers from B) annotated vs automated. D) Bar plot of the track similarity to manually annotated (Annot.) branch tracks. To compare performance against random assignment, the same number of tracks was sampled n times from the dense tracks and from all assigned tracks of the organoid. E) Bar plot of the temporal error measured by the average frame difference between the manually annotated (Annot.) and predicted duration of the organoid track (Auto). The same sampling strategy as D) was used to assess the performance against random assignment.

Method	Branch Counts	Track Similarity	Temporal Difference
Annot.(manual)	13.5±2.3	1	0
Auto	17.5±2.9	**0.55±0.04**	**130±17** frames
Rand. (Dense)	-	0.37± 0.01	154±11 frames
Rand. (Assigned)	-	0.37±0.01	153±11 frames

Table 6.1: Summary of organoid branch tracking validation. Reported errors are the mean±S.E.M, (standard error of the mean) over the 8 organoid videos.

MOSES mesh formulation then tends to overestimate the number of branches. Preliminary experiments with increasing the tracking density by introducing new superpixels over time in a similar manner to dense trajectories (Wang et al. 2013), c.f. Ch.7.3.1 suggests a significant increase in accuracy is possible. For track similarity and temporal accuracy, (Fig.6.10D,E) the performance of the proposed branch tracking algorithms consistently outperforms random assignment of tracks (used here as a baseline algorithm), scoring higher in track similarity and lower in temporal differences. To randomly assign tracks we pick the number of tracks equal to the number of manually annotated tracks with a uniform prior from i) the set of all superpixel tracks assigned to the organoid (denoted 'Rand_track') or ii) the reduced set of 'densest tracks' (denoted 'Rand_Dense')), where each track has a uniform chance of being picked. The 'picking' is carried out n number of times akin to Monte Carlo simulation to compute mean scores and confidence intervals. Table.6.1 summarises the mean performance where (Assigned) and (Dense) is used to refer to the aforementioned random assignment procedure i) and ii) respectively. It should be noted that the majority of the temporal difference is due to the inaccuracy in identifying the start of branching which is very subtle and difficult even for a human annotator to define. For the majority of the 468 frames the most motion due to branch growth only occurs in the last 1/3rd of all frames therefore an average frame difference of 130 frames is acceptable. In most cases, the algorithm performs under this mean error, (Fig.6.10E).

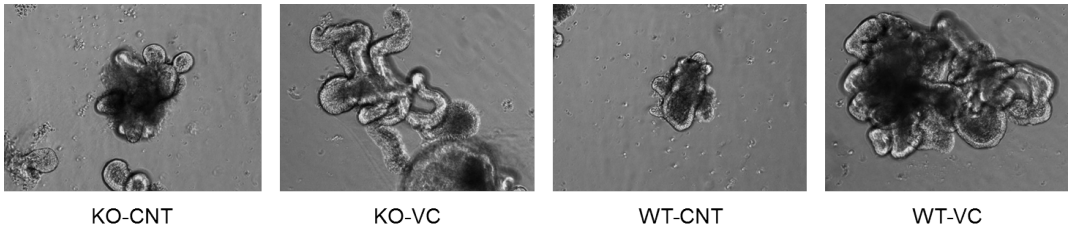

Figure 6.11: Phenotype effect of VC treatment to intestinal organoids. Static video snapshots of the final frame, frame 468. One example is shown for each condition.

6.7 Motion Analysis of Intestinal Organoids with Treatment

The intestinal organoids exhibit striking phenotypic changes after VC treatment and is ideal to test the developed algorithm for morphological analysis. Wild type (WT) and P53$^{-/-}$ (abbreviated as KO for knockout) mouse intestinal organoids were subject to VC (valproic acid + CHIR99021) treatment (Yin et al. 2014) and filmed under timelapse for 7 days (Ch.6.4). Qualitatively, this treatment induced a greater amount of branching and result in larger organoids. Visually, WT and KO organoids were very similar, (Fig.6.11). For representative quantitative analysis of organoid growth dynamics the automatically extracted organoids was filtered as described (Ch.6.5.4) to extract only organoids whose motion is primarily due to growth. This led to a reduction from a total of 223 organoids detected from the 80 videos to 89 organoids for analysis.

VC Treatment Increases Motion, Growth and Branching

When the organoid motion is primarily due to growth, the MOSES strain curve reflects the organoid shape as it evolves in time and gives rise to a characteristic growth curve which exhibits fast exponential growth followed by a plateau to stability, (Fig.6.12A). For quantitative analysis we compared the maximum strain (reflective of the maximum organoid size) of the MOSES strain curve for organoids from the four conditions. The results agree with qualitative observations (Fig.6.11). There is little difference in final size between WT and KO organoids but there is clear increased size following VC treatment in both cases. Additionally we find that

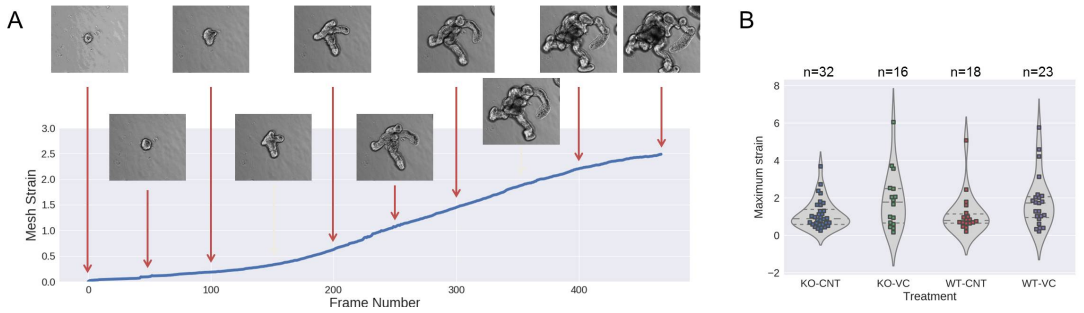

Figure 6.12: Valproic acid intestinal organoid motion analysis. A) MOSES strain curve for an organoid with static video snapshots at the indicated frame number. B) Violin and swarmplot of maximum (max.) strain for organoids of each treatment condition. Each point is an organoid. Width of the plot indicates the number of organoids at this strain value. Lines inside each violin indicate the quartiles of the data.

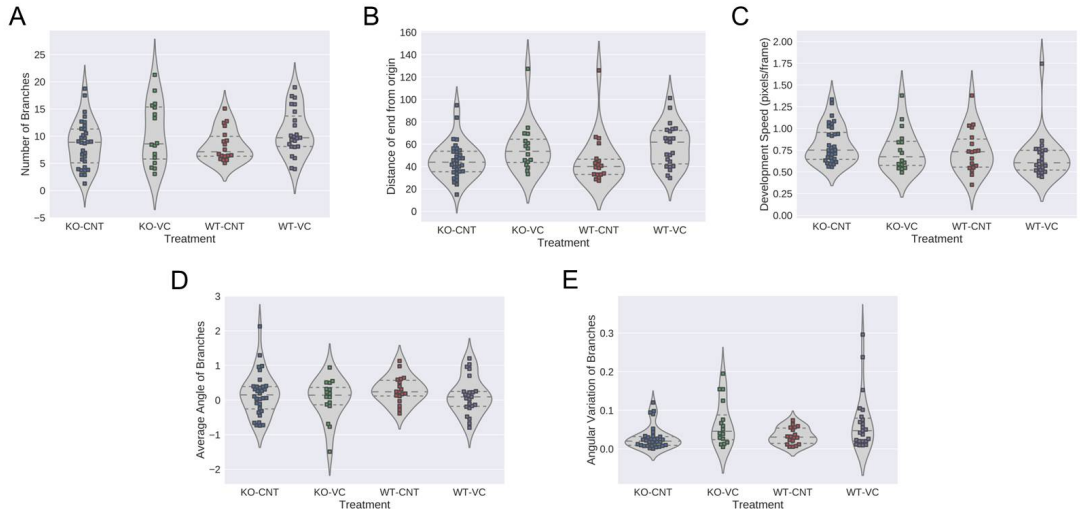

Figure 6.13: Valproic acid intestinal organoid branch analysis. A) Number of branches. B) Distance between the end and start position of the average organoid branch. C) The number of pixels moved per frame of the average organoid branch. D) Average orientation angle of the organoid branches. E) Angular variation (variance) of the organoid branches.

the treatment effect is more heterogeneous in the case of KO organoids where there appears to be the emergence of a bimodal size distribution, (Fig.6.12B).

We next analysed the branching dynamics. In general the effect is heterogeneous reflecting the heterogeneity of the organoid culture. We find organoid subpopulations with increased number of branches after VC treatment, (Fig.6.13A), generally longer directional growth of organoids evidenced by an increased distance from the end position to the start position and therefore more elongated branches, (Fig.6.13B)

at a slight decrease in development speed, (Fig.6.13C). An averaged branching angle of 0 across all conditions suggest general outward growth of branches as expected. Increased angular variation considering an average branching angle of 0 suggests slightly enriched directional growth in a subpopulation of organoids following VC treatment. With respect to the branching dynamics we find slight differences between WT and KO organoids mainly due to increased number of branches in KO, (Fig.6.13A) and faster branch movement, (Fig.6.13C). In conclusion VC treatment strongly perturbs intestinal organoid branching dynamics and results in larger organoids and enhanced branching.

RNA-Seq Results

An independent set of organoids were transcriptome profiled using RNA-seq. Each sample represents one well of a multi-well plate with each well containing 10's-100's of organoids. An unbiased analysis of the sequencing reads using dimensional reduction with multidimensional scaling of the log counts per million ($\log_2$(CPM)) confirms that VC treatment dominates any genetic differences in organoid (i.e. WT vs KO, Fig.6.14A). Kegg pathway enrichment analysis of differentially expressed genes identified using EdgeR (Robinson et al. 2010) further finds 'the regulation of actin cytoskeleton', essential in motility, cell proliferation and shape changes as the pathway with the most number of genes, (Fig.6.14B). This is followed by 'axon guidance' which may point to the increased elongated branching dynamics observed following VC treatment. Many of the differentially expressed actin cytoskeleton genes play a role in motion, (Fig.6.14C) but the exact nature is difficult to interpret with a mixture of up and downregulation, (Fig.6.14D). From Fig.6.14D, valproic acid as a HDAC inhibitor in the VC treatment broadly affects large sections of the genome and thus a dramatic phenotypic change is expected. To summarise, the independent RNA-seq results confirm the automated MOSES video analysis of organoids.

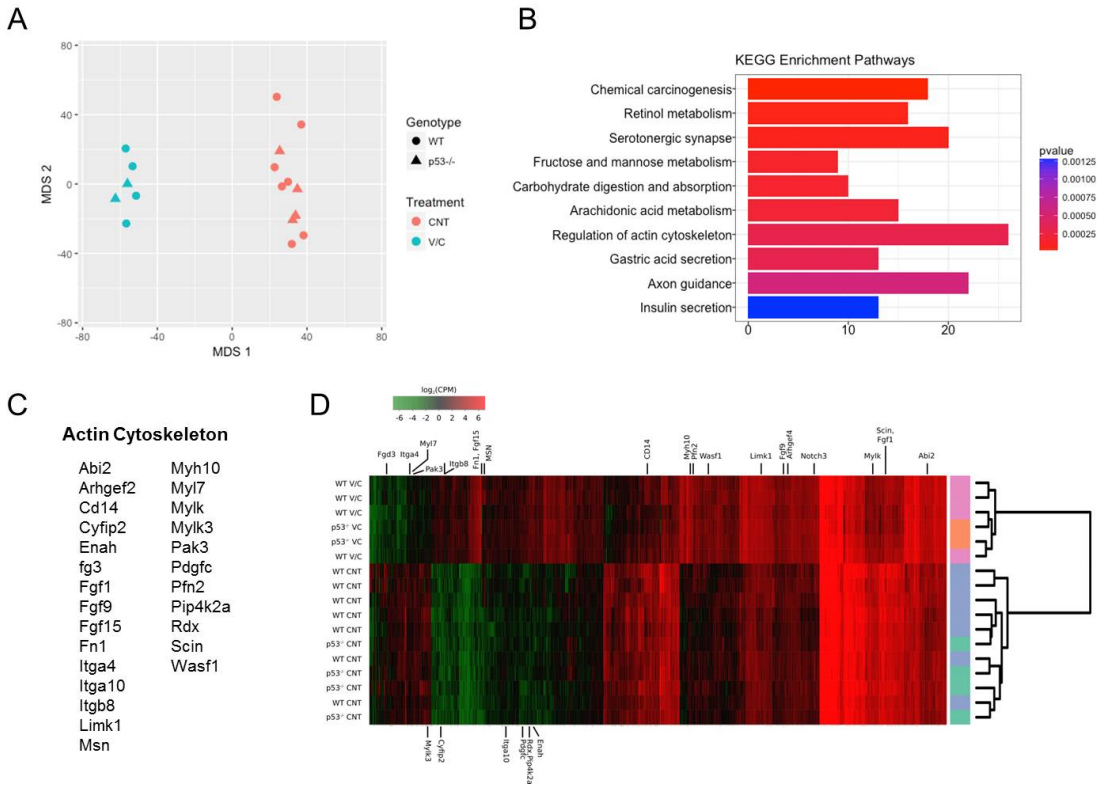

Figure 6.14: RNA-seq assessment of VC treatment on intestinal organoids. A) Two component multi-dimensional scaling (MDS) for dimensional reduction based on the log-fold counts per million $\log_2$(CPM) of RNA-seq samples. B) Statistically significant KEGG enriched pathways based on differentially expressed genes ($p \leq 0.05, \log_2(\text{FC}) \geq 1$) showing the number of genes in each pathway. C) Genes under the actin cytoskeleton pathway of B). D) Heatmap of RNA-seq samples with $\log_2$(CPM) of differentially expressed genes. Actin cytoskeleton pathway genes are labelled. Hierarchical clustering uses the Ward method.

The analysis of the full intestinal organoid dataset, 80 videos in total, each video on average 512x512 pixel resolution, 468 frames, 5000 superpixels can be completed in $<$ 1 day. For 96-well plate screens, one video per well with $\sim$60 organoids, 2048x2048 pixel resolution, 120 frames and 20000 superpixels analysis can be completed in $<$ 3 days. These timings are reported on a single machine (3.3GHz, 64GB RAM). MOSES is therefore capable of high-throughput analysis of organoid timelapse videos.

6.8 Summary and Conclusions

- Organoids have emerged as a promising *in-vitro* model that could be grown using primary patient biopsies as a surrogate for personalised treatment. Current screening methods are primarily based on viability/toxicity, (alive or dead) followed by sequencing approaches. They do not consider organoid size and morphology which can be both organ and disease-specific. We hypothesize quantitative timelapse imaging could provide non-destructive, higher-throughput, cost-effective dynamic phenotyping for more targeted downstream sequencing.
- Extending the MOSES pipeline introduced for analysing epithelial sheet interactions we demonstrate a simple, minimal parameter framework to analyse organoid and organoid-like objects imaged with label-free microscopy. We show
    1. Spatiotemporal motion segmentation of individual organoids even in the presence of multiple organoids.
        - Single organoid detection – 100% accurate
        - Multiple organoid detection (1-8) – 73% accurate
    2. Spatiotemporal tracking of organoid morphology
        - Recovery of boundaries for small organoids and bounding box for larger organoids.
        - Bounding box continuous tracking of organoid location and approximate shape.
    3. Spatiotemporal tracking of organoid branching
        - Branch detection accuracy of 70.8%, high correlation (0.93) to manual counting.
        - Track similarity score of 0.55 to manually annotated tracks.
- Automated morphological analysis is in concordance with independent bulk RNA-seq of organoids. In addition the analysis shows quantification of individual organoid heterogeneity from the same condition thus allowing targeted specific subpopulation analysis.

7 Discussion and Future Directions

Contents

7.1 Overview

The aim of this thesis was to formulate a simple, potentially generally applicable approach for extracting and analysing motion patterns in a wide range of biological videos for comprehensive spatiotemporal phenotyping. In particular the framework should be sufficiently flexible to accomplish a wide variety of general computer vision tasks; i) object segmentation, ii) object tracking, iii) video classification. It should not utilise segmentation-based tracking or automated cluster determination techniques which are not robust and require a high degree of fine-tuning. Importantly a unique motion signature should be able to be derived for each video for motion

similarity comparison and suitable for integration with additional datasets to enable deep phenotyping for precision medicine as motivated in Ch.1:General Introduction.

In Ch.2: Extracting and Describing Motion, dense optical flow was shown to be an excellent agnostic method for extracting motion from image sequences and that the best hand-crafted video classification method exploited the construction of dense trajectories from the extracted dense optical flow. In Ch.3: Models for Quantitative Motion Analysis, the Lagrangian approach in particular the theory of Finite Time Lyapunov Exponents (FTLE) motivates the forward and backward construction of dense trajectories theoretically as the discovery of local motion 'sinks' and 'sources'. We also saw how graphs could potentially provide a unifying approach for analysing local to global motion, by providing a structure to agglomerate information. Successful proposed collectiveness measures and group-level scenario recognition models all utilise some notion of a graph.

Based on this literature review, our proposed framework which we call Motion SEnsing Superpixels (MOSES) brings together the core concepts of dense optical flow trajectories and construction of graphs with a mid-level visual representation provided by superpixels. Compared to similar literature attempts, (e.g. Fragkiadaki et al. 2012) our approach is refreshingly simple and back to basics; i) agnostic equipartitioning of the image into smaller regions bigger than a pixel (superpixels), ii) the number of superpixels once generated are fixed and are propagated according to dense optical flow for the full video length (unlike other dense trajectory methods, (Brox and Malik 2010,Wang et al. 2013)) and iii) resulting trajectories are deterministically linked based on a notion of distance to form dynamic meshes (unlike the multiscale path similarity in Zhou et al. 2013). In Ch.4: Motion Sensing Superpixels, we demonstrated and explored the potential of MOSES using four different datasets and established the general notion of motion signatures and motion saliency maps within the MOSES framework for exploratory video analysis. Motion signatures allowed us to interrogate the motion heterogeneity between videos whilst

motion saliency provided visual spatial localisation of important motion events such as the fingers of a piano player and spatial areas of attraction for neutrophils. In Ch.5: Phenotyping Cell Population Interactions, we analysed in-depth the interactions between two epithelial sheets with cell lines constituting cells at the squamous-columnar junction under fluorescence microscopy and demonstrated the power of MOSES to derive robust metrics for quantifying boundary formation in the presence of external perturbation on a continuous scale and the ability to construct predictive motion maps with MOSES motion signatures. Finally in Ch.6: Organoids, we analysed in-depth the spatiotemporal development of intestinal organoids with MOSES filmed under label-free phase contrast microscopy. Using the dynamic meshes as a basis we show the ability to recover spatiotemporally the individual motion of each organoid even when multiple are present in the field-of-view and may overlap as they grow. From this information we showed the ability to recover the organoid shape (motion saliency map) and the ability to localize bounding boxes and remarkably the organoid branching. Comparison of the automated video analysis results of organoids grown under different conditions showed strong agreement with an independent RNA-squenced cohort verifying the potential for complementary integration of imaging with sequencing approaches.

In conclusion MOSES is a powerful conceptually simple modality-independent analysis framework. Through extensive experimentation we have shown that it is robust and applicable to high-content analysis. It is exceptionally useful where the scientific question of interest does not necessarily require accurate object boundary information provided by more complex approaches such as single-cell tracking and image segmentation. In addition the framework is modular and can be combined with other approaches, for example for single cell tracking the dense optical flow would not be recommended. Instead the dynamic meshes can be built upon the resultant tracks of an appropriate specialised single-cell tracker. Despite its successes MOSES still has limitations with respect to its current implementation that can be improved for example generalisation of the graph-based superpixel track assignment,

devising a less data specific way for initial object detection and extension of the implemented 2D analysis to 3D timelapse datasets. The latter extension in principle is relatively straightforward using supervoxels instead of superpixels and higher dimensional optical flow. In practice, the z depth spatial resolution of microscopes is often worse than the x, y resolution and a weighting parameter may need to be incorporated and explored in future work.

The validation of the developed MOSES methodology has been non-trivial. A more rigorous and objective evaluation instead of 'visual' quality assessment as pursued in computer vision or CT/MRI imaging datasets is desirable for the thesis but with the microscopy datasets presented here it is difficult to establish the necessary ground-truth. Large, annotated datasets are scarce (e.g. biological developmental videos), do not contain the necessary annotation (e.g. two-cell population migration dataset with boundary formation or collective motion annotations) or non-existent (organoids timelapse). As such, given the time constraints we have instead pursued in this thesis 'experimental validation' where the algorithm prediction and analysis output is compared to the results of biological experiments with known expected outcomes e.g. branching of intestinal organoids. Having now established a sufficiently broad, general and robust pipeline, future work will seek more objective validation of MOSES using the existing implementation to semi-automate the creation of large annotated datasets.

In the remainder of this chapter we first discuss how to complete the work on organoids by application of the automated MOSES pipeline to analyse a screen before describing some of the major algorithmic improvements that can be explored to extend the robustness and capabilities of MOSES.

7.2 Automated Organoid Screening

The primary utility of an automated organoid analysis system is in the high-content screening of drugs for personalised treatment recommendation. As a first

step, to show proof-of-principle, in the lab we have collected primary human organoids from routine endoscopy in collaboration with the hospital. We intend to subject these normal organoids to commonly prescribed chemotherapy drugs such as 5-fluorouracil, cisplatin, cetuximab and monitor their development under timelapse in a toxicity screen with 96 well plates. The ideal chemotherapy drug should maximally affect cancer cells whilst minimising the effect on normal cells. Tumour organoids are more difficult to acquire and is an ongoing process. To conclude the organoid component of this thesis, the proposed pilot experiment with normal organoids would in principle demonstrate the ability to acquire timelapse acquisitions of multiple organoids from each well in a 96-well plate and reveal any potential complementary benefits of an additional in-depth morphological analysis to the standard live/death viability assay. In a second screening, we could co-culture tumour and normal organoids together where the developed automated analysis will provide an unparalleled extra dimension in monitoring and assessing potential spatiotemporal interactions between the two organoid types. Naturally this interaction analysis is applicable to any two or multispecies interaction. Another avenue that can be explored using co-cultures for example is the gut-brain axis whose cross-talk is increasingly becoming recognised as being critical in the proper maintenance of gastrointestinal homeostasis and higher cognitive functions, (Foster and Neufeld 2013; Clarke et al. 2013; Carabotti et al. 2015).

7.3 Algorithm Improvements to MOSES

There are several key limitations to MOSES which can be addressed for increased performance and general applicability. First, from our experiments on UCF-101, a standard human action recognition benchmark (Ch.4.6.1) the proposed 1D motion signatures do not sufficiently capture appearance which for general videos is critical as demonstrated by the improved dense trajectories approach of Wang and Schmid 2013 which extracts numerous appearance features in addition to motion-based features. Second, from the single cell tracking experiments (Ch.4.6.2) the current method of motion analysis does not assign tracks to new objects entering into the

field of view or factor into account different contributions to motion such as cell division. Third, from the long-time tracking of *drosophila* embryo development (Ch.4.6.4) and organoid shape inference from trajectories, (Ch.6.5.4) for videos with highly dynamic movement, the original densely seeded superpixels tend to aggregate and sample the movement in the same region over time. Having fixed the number of superpixels *a priori*, this leads to insufficient sampling of the spatial region in later frames. For the *drosophila* embryo development the later occurring motion regimes became less pronounced. For organoids there was difficulty in recovering the boundary shape for large organoids at later times. Below we address possible solutions to each of the outlined problems in order of increasing difficulty based on extending the current formulation of MOSES.

7.3.1 Improving Density of Tracking

To improve the density of tracking, additional superpixels for tracking must be added. There are two approaches. The easiest method is to divide long videos into separate clips and then to run MOSES on the separated clips. Analysis such as motion segmentation can be extracted for each clip and the results concatenated. One drawback of this approach is that it may forfeit the benefits of temporal continuity. In organoids the temporal continuity proved crucial to separate those that grow into each other. Had the video been divided into video clips the accuracy would have been reduced. At later frames, image segmentation due to crowding becomes much more difficult and harder to seed the instance segmentation required for the initial frame. The second approach is to follow the existing literature (Brox and Malik 2010; Fragkiadaki et al. 2012) to continuously detect sparse regions and seed new superpixels. The drawback of this approach is that it will result in a varied number of trajectories per video with variable temporal lengths. We would need to adapt some of the analysis to fit this for example resultant trajectories may require temporal padding and the MOSES mesh definition based on a static neighbourhood definition would need to be refined.

7.3.2 Improving Motion Signatures

The proposed motion signatures such as the forward MOSES signature also known as the MOSES mesh strain or backward MOSES signature in Ch.4.4.2 collapse out the spatial component of the dynamic mesh motion features by averaging over the superpixels. The averaging operation can be seen as a type of 'pooling' operation. One way to recover this information in a manner still suitable for downstream algorithms such as PCA is to apply bag-of-words or fisher vector encoding (c.f. Ch.2.3.1) given a video collection. A second method is to use recurrent neural network (RNN) type units such as LSTM that can directly operate on spatiotemporal features. An encoder-decoder framework could be used to compress the features into a feature vector or an LSTM network used directly for classification. To include appearance based features we can follow similar approaches as those used for action recognition, (Ch.2.3.1, 2.3.2) extracting for each superpixel along its trajectory handcrafted features such as HoG or CNN features using a pretrained CNN. The proposed extensions can be verified on the UCF-101 dataset and on the two cell epithelial population dataset to see if it leads to improved motion map learning. A potential drawback of the above approaches is the inability to learn a mesh connecting superpixels. Below, (Ch.7.4) we discuss the possibility of using convolutional graph networks which extends deep learning to graph-structured data that has emerged in recent years.

7.3.3 Deconvolution of Motion Contributions

There can be many different contributions to the measured global motion from optical flow due to cellular processes for example in addition to migration there is cell division with the splitting motion of cells and the subsequent introduction of a new 'object' capable of independent motion. There is also cell shape changes which do not affect the centroid position of the cell and cell death which leads to an object disappearing. Additionally there may be camera motion such as camera shake and zoom which can induce apparent motion artifacts. Camera motion in particular presents problems for *in-vivo* samples which cannot be readily fixed c.f.

intra-vital imaging (Kedrin et al. 2008; Pittet and Weissleder 2011). In general it is very difficult to automatically separate the different motion contributions without explicit measurement of each source independently. For camera motion, the restricted set of movements of the camera can be modelled and its contribution removed using registration techniques such as we did with organoids (Ch.6.5.2) or through explicit homography estimation (Wang and Schmid 2013). For processes such as cell division this is solved with image segmentation and tracking. But what if segmentation is not possible as with the confluent 2D epithelial sheet? Progress might be made based on targeting the different motion timescales that characterise certain cellular processes. Migration for example tends to be a directional movement and can occur in culture continuously through time whereas cell division only occurs periodically at set time-intervals. We are inspired in this thought by the work of Angelini et al. 2011 whom applied spatiotemporal Fourier transform techniques to infer the average division time of cells in a confluent monolayer (c.f Ch.3.3.2), the use of the Fourier-transformed velocity autocorrelation function to isolate different components of the motion in Li et al. 2011 and the work of Koslover et al. 2016 whom apply wavelet methods to isolate the stochastic motion of lysosomes within motile migrating neutrophil-like cells. An alternative temporally oriented approach along these lines is to decompose the motion trajectory into 'states', temporal segments which exhibit similar statistics using a non-parametric method such as Hidden Markov models (HMM) or the Sticky HDP-HMM (Fox et al. 2007). The states may later be associated to particular cellular processes. In Held et al. 2010 this idea was exploited using HMMs to discover the appearance of cells at different stages of the cell cycle. In summary much more research is required to find a general method. If cells are sparse, it is probably best to utilise a single cell tracker.

7.4 Possible Extensions for End-to-end learning

Many important real-world datasets come in the form of graphs or networks: social networks, knowledge graphs, protein-interaction networks, the World Wide Web to mention just a few. Sound and images too can also be viewed as regular

graphs, (Fig.7.1A). An image for example is a regular grid-like graph built on pixels, one of the key ideas behind Markov Random Fields (Blake et al. 2011). Graph convolutional networks (GCNs) have recently emerged as a set of methods to generalize the locally shared filters key to the success of convolutional neural networks for images to structured datasets. As a recent field, various formulations have been proposed such as problem-specific specialized architectures (Duvenaud et al. 2015; Li et al. 2015; Jain et al. 2016a), extensions of graph convolutions known from spectral theory (Bruna et al. 2013; Henaff et al. 2015) to define parameterized filters that are used in a multi-layer neural network model and works that focus on bridging the gap between the fast heuristics of the former and the slower, but somewhat more principled spectral approach of the latter (Defferrard et al. 2016; Kipf and Welling 2016). We give a brief exposition of GCNs before discussing how this might provide a principled approach for extending MOSES to take advantage of the benefits of data-driven learning when large datasets are available. Thomas Kipf provides an excellent writeup of GCNs, (Kipf 2016).

7.4.1 Graph Convolutional Networks (GCNs)

For graph convolutional networks (GCNs), the goal is to learn a function of features on a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$. In general GCNs have the following form which takes the inputs below to give an output network prediction.

GCN Input

- A feature description x_i for every node i summarized in a $N \times D$ feature matrix X (N: number of nodes, D: number of input features).

- A description of the graph structure in matrix form; typically in the form of an adjacency matrix A.

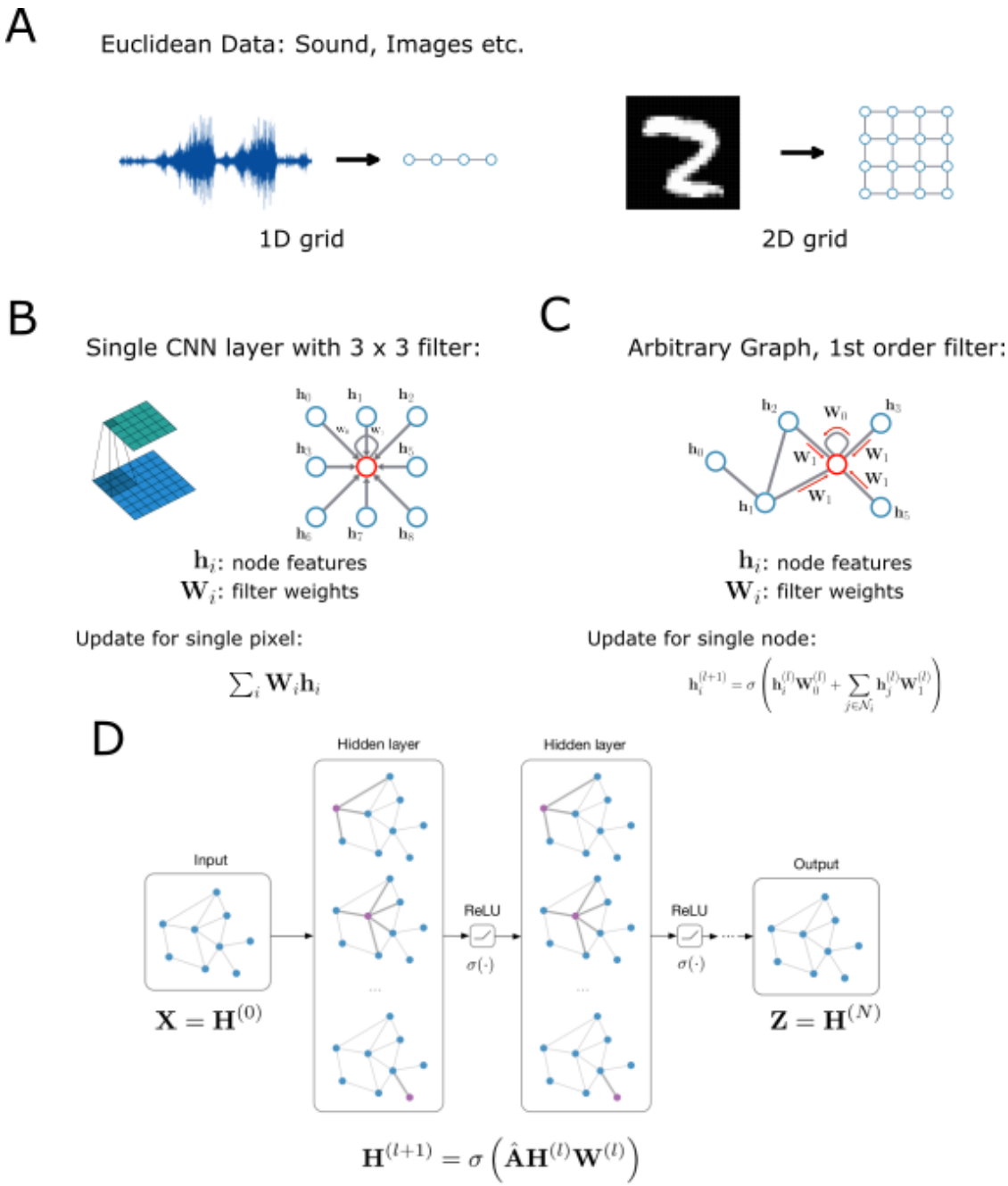

Figure 7.1: Graph convolutional networks. A) Sequential data such as sound can be viewed as a 1D grid graph and images viewed as a 2D grid graph. B) Illustration of the update rule for a single pixel in convolutional neural networks. C) Analogous update rule for a single node of an arbitrary graph with graph convolutional networks. D) Illustration of the multilayer application of the update rule in C). Adapted from Kipf and Welling 2016.

GCN output

- Node-level output, Z ($N \times F$ feature matrix, F: number of output features per node).

Every neural network layer can then be written as a non-linear function just like with a multilayer neural network.

$$H^{(l+1)} = f(H^{(l)}, A) \tag{7.1}$$

with $H^{(0)} = X$, and $H^{(L)} = Z$, L is the number of layers. Specific models then differ only in how $f(\cdot, \cdot)$ is chosen and parameterized.

A Simple GCN In analogy with convolutional neural networks, (Fig.7.1B) consider the graph in Fig.7.1C with 1st-order message passing and let $f(\cdot)$ be a weighted linear update followed by a non-linear activation function, σ like ReLU, then:

$$\mathbf{h}_i^{l+1} = \sigma\left(\mathbf{h}_i^l W_0^{(l)} + \sum_{j \in \mathcal{N}_i} \mathbf{h}_j^l W_1^{(l)}\right)$$

where $\mathcal{N}_i$ is the neighbour indices. In general we can write

$$f(H^{(l)}, A) = \sigma(\hat{A} H^{(l)} W^{(l)}) \tag{7.2}$$

where $\hat{A} = A + I$ is the adjacency matrix with self-loops, $W^{(l)}$ is the weight matrix in the l-th neural network layer. $\hat{A}$ should be normalized such that multiplication with $\hat{A}$ will not change the scale of the feature vectors. Normalizing $\hat{A}$ such that all rows sum to one is equivalent to $\hat{D}^{-1}\hat{A}$ where $\hat{D}$ is the diagonal node degree matrix of $\hat{A}$. Multiplying with $\hat{D}^{-1}\hat{A}$ corresponds to taking the average of neighboring node features (c.f. MOSES dynamic mesh signatures). In practice, dynamics is more interesting if a symmetric normalization, $\hat{D}^{-1}\hat{A}\hat{D}^{-1}$ is used which no longer amounts to an averaging of neighboring nodes. This is the propagation rule introduced in Kipf and Welling 2016:

$$f(H^{(l)}, A) = \sigma(\hat{D}^{-\frac{1}{2}}\hat{A}\hat{D}^{-\frac{1}{2}} H^{(l)} W^{(l)}) \tag{7.3}$$

and the local weights can be trained using gradient descent, (Fig.7.1D). It is however important to note that this formulation of graph convolutional networks applied to grid data such as images does not yield the same representational power as an equivalent convolutional neural network, (Huszár 2016) where it reduces to a centre-surround pattern. This is also graphically evident when we compare the two update rules, (Fig.7.1B,C).

Nevertheless this formulation is ideal for MOSES and critically takes advantage of the structured representation provided by the dynamic meshes which normal convolutional networks cannot. Recall that the dynamic mesh proposed in this thesis is generally defined by T feature matrices of dimension, $N \times N_{features}$ (Ch.4.4.2) together with corresponding T adjacency matrices of dimension, $N \times N$ where N is the number of superpixels, T is the total number of video frames and $N_{features}$ is the number of input features. This is effectively a sequence of graphs. The GCN presented above is formulated for a single static graph. We require an additional extension to the temporal dimension in order to provide a means of generating a compact motion signature that summarises the full motion content and to incorporate dynamically changing edge connections and dynamic vertex features, (Manessi et al. 2017). However this static formulation is already extremely useful. In particular it may provide a systematic way to fuse spatial graphs for which there is limited literature treatment. Most graph fusion approaches assume non-spatial graphs, graphs whose nodes are not anchored to a particular (x, y) position. Recall in chapter 5 we showed the extraction of two meshes, one for each colour representing the different cell types. To compute motion metrics we independently extracted the measurements from both meshes then averaged the result. A more systematic approach is to fuse the two meshes and extract the motion parameters. Using GCNs we could train a two-input, one-output GCN where the inputs are the two meshes and the output is the fused mesh at the same time point. The GCN should be trained to minimise the regression error over all time instances. For training, the 'ideal' fused mesh would be the mesh extracted

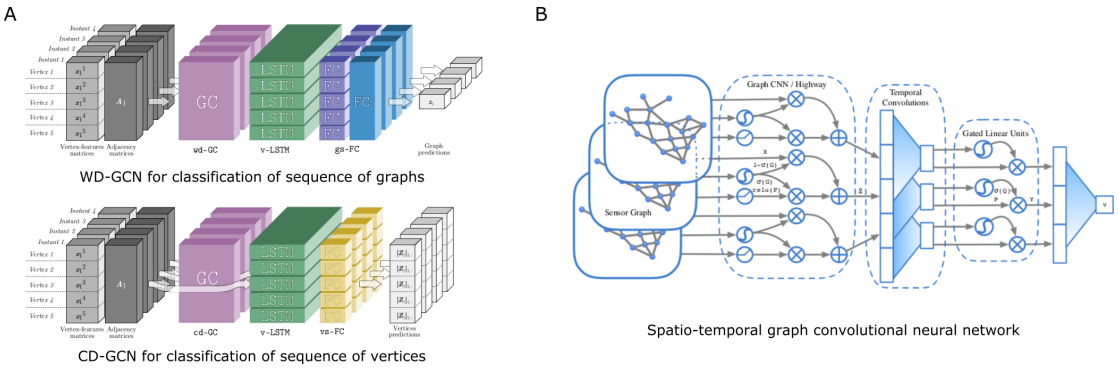

Figure 7.2: Spatiotemporal graph convolutional networks. A) Two models proposed by Manessi et al. 2017 that extend the graph convolutional networks of Kipf and Welling 2016 with LSTM for spatiotemporal classification (sequence of graphs) and regresion (sequence of vertices). Adapted from Manessi et al. 2017. B) The spatiotemporal graph convolutional network of Yu et al. 2017 that employs temporal convolutions and gated linear units for prediction. Adapted from Yu et al. 2017.

by treating the fluorescent videos as a grayscale image sequence. At present, we have only used heuristic approaches to visualise fused meshes.

7.4.2 Spatiotemporal GCNs

In 2017, a few papers have extended the general GCN formulation to temporal structured data with promising results. Manessi et al. 2017 propose two models, (Fig.7.2A) one for classification (sequence of graphs) and one for regression (sequence of vertices) that extend GCNs in a similar manner to Conv-LSTMs for images (Ch.2.3.2). The key idea is to essentially use multiple instances of graph convolutional layers with shared weights to extract graph features which are then fed to LSTM units that subsequently exploit the sequential information. Fully connected layers are finally used to hierarchically pool the sequential features to give a sequence of numbers for classification or a sequence of vertices if the output is desired to be a graph. The authors show through statistical testing the performance of this architecture was superior than baseline architectures that utilise only graph features with no temporal knowledge or temporal features with no knowledge of the graph structure. Instead of using recurrent neural networks which are more difficult and slower to train, Yu et al. 2017 chose to exploit convolutional sequence learning which uses convolutional neural networks for more stable training and

faster convergence, (Fig.7.2). In doing so, Yu et al. 2017 demonstrated state-of-the-art performance on real collected traffic flow data. Research in this area is still however very young. Many factors such as the number of layers, choice of propagation rule $f(\cdot)$ and issues of data overfitting and performance with increased number of nodes and vertices remain to be explored. Nevertheless these initial works appear promising. We propose to build upon and extend the developed architectures to leverage the developed dynamic meshes for feature learning and classification. As a final remark we note the numerous research possibilities that could exist by combining image CNNs with GCNs for example one could ask whether we could directly learn to extract the graph structure and learn the salient spatial sampling points (superpixels) direct from the input sequence of images by combining a CNN extractor that proposes graphs fed into a spatiotemporal GCN and trained in an end-to-end manner.

References

Achanta, Radhakrishna et al. (2008). "Salient Region Detection and Segmentation". In: *Computer Vision Systems, 6th International Conference, ICVS 2008, Santorini, Greece, May 12-15, 2008, Proceedings*. Ed. by Antonios Gasteratos, Markus Vincze, and John K. Tsotsos. Vol. 5008. Lecture Notes in Computer Science. Springer, pp. 66–75.

Achanta, Radhakrishna et al. (2012). "SLIC superpixels compared to state-of-the-art superpixel methods". In: *IEEE transactions on pattern analysis and machine intelligence* 34.11, pp. 2274–2282.

Al-Kasspooles, Mazin et al. (1993). "Amplification and overexpression of the EGFR and erbBB2 genes in human esophageal adenocarcinomas". In: *International journal of cancer* 54.2, pp. 213–219.

Alberts, Bruce (2017). *Molecular biology of the cell*. Garland science.

Alexe, Bogdan, Thomas Deselaers, and Vittorio Ferrari (2012). "Measuring the objectness of image windows". In: *IEEE transactions on pattern analysis and machine intelligence* 34.11, pp. 2189–2202.

Ali, Saad and Mubarak Shah (2007). "A lagrangian particle dynamics approach for crowd flow segmentation and stability analysis". In: *Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE Conference on*. IEEE, pp. 1–6.

Ali, Saad, Arslan Basharat, and Mubarak Shah (2007). "Chaotic invariants for human action recognition". In: *Computer Vision, 2007. ICCV 2007. IEEE 11th International Conference on*. IEEE, pp. 1–8.

Alt, Silvanus, Poulami Ganguly, and Guillaume Salbreux (2017a). "Vertex models: from cell mechanics to tissue morphogenesis." In: *Philosophical transactions of the Royal Society of London. Series B, Biological sciences* 372 (1720).

— (2017b). "Vertex models: from cell mechanics to tissue morphogenesis". In: *Phil. Trans. R. Soc. B* 372.1720, p. 20150520.

Amat, Fernando et al. (2014). "Fast, accurate reconstruction of cell lineages from large-scale fluorescence microscopy data". In: *Nature methods*.

Angelini, Thomas E et al. (2010). "Cell migration driven by cooperative substrate deformation patterns". In: *Physical review letters* 104.16, p. 168104.

Angelini, Thomas E et al. (2011). "Glass-like dynamics of collective cell migration." In: *Proceedings of the National Academy of Sciences of the United States of America* 108 (12), pp. 4714–4719.

Arbelaez, Pablo (2006). "Boundary extraction in natural images using ultrametric contour maps". In: *Computer Vision and Pattern Recognition Workshop, 2006. CVPRW'06. Conference on*. IEEE, pp. 182–182.

Arbeláez, Pablo et al. (2014). "Multiscale combinatorial grouping". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 328–335.

Avramov, I (1998). "Viscosity of glassforming melts". In: *Journal of non-crystalline solids* 238.1, pp. 6–10.

Badrinarayanan, Vijay, Alex Kendall, and Roberto Cipolla (2015). "Segnet: A deep convolutional encoder-decoder architecture for image segmentation". In: *arXiv preprint arXiv:1511.00561*.

Bao, Zhirong et al. (2006). "Automated cell lineage tracing in Caenorhabditis elegans". In: *Proceedings of the National Academy of Sciences of the United States of America* 103.8, pp. 2707–2712.

Barker, Bethan L and Christopher E Brightling (2013). "Phenotyping the heterogeneity of chronic obstructive pulmonary disease". In: *Clinical Science* 124.6, pp. 371–387.

Basset, Antoine et al. (2017). "An extended model of vesicle fusion at the plasma membrane to estimate protein lateral diffusion from TIRF microscopy images". In: *BMC bioinformatics* 18.1, p. 352.

Behl, Aseem et al. (2017). "Bounding Boxes, Segmentations and Object Coordinates: How Important is Recognition for 3D Scene Flow Estimation in Autonomous Driving Scenarios?" In: *IEEE International Conference on Computer Vision (ICCV), 2017*.

Bershteyn, Marina et al. (2017). "Human iPSC-derived cerebral organoids model cellular features of lissencephaly and reveal prolonged mitosis of outer radial glia". In: *Cell stem cell* 20.4, pp. 435–449.

Bialek, William et al. (2012). "Statistical mechanics for natural flocks of birds". In: *Proceedings of the National Academy of Sciences* 109.13, pp. 4786–4791.

Bilen, Hakan et al. (2016a). "Action Recognition with Dynamic Image Networks". In: arXiv: `1612.00738v1 [cs.CV]`.

Bilen, Hakan et al. (2016b). "Dynamic image networks for action recognition". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3034–3042.

Bise, Ryoma, Zhaozheng Yin, and Takeo Kanade (2011). "Reliable cell tracking by global data association". In: *Biomedical Imaging: From Nano to Macro, 2011 IEEE International Symposium on*. IEEE, pp. 1004–1010.

Blake, Andrew, Pushmeet Kohli, and Carsten Rother (2011). *Markov random fields for vision and image processing*. Mit Press.

Boehnke, Karsten et al. (2016). "Assay Establishment and Validation of a High-Throughput Screening Platform for Three-Dimensional Patient-Derived Colon Cancer Organoid Cultures". In: *Journal of biomolecular screening* 21.9, pp. 931–941.

Bonneau, Stéphane, Maxime Dahan, and Laurent D Cohen (2005). "Single quantum dot tracking based on perceptual grouping using minimal paths in a spatiotemporal volume". In: *IEEE transactions on image processing* 14.9, pp. 1384–1395.

Boonstra, Jurjen J et al. (2010). "Verification and unmasking of widely used human esophageal adenocarcinoma cell lines". In: *Journal of the National Cancer Institute* 102.4, pp. 271–274.

Boquet-Pujadas, Aleix et al. (2017). "BioFlow: a non-invasive, image-based method to measure speed, pressure and forces inside living cells". In: *Scientific reports* 7.1, p. 9178.

Bouguet, Jean-Yves (2001). "Pyramidal implementation of the affine lucas kanade feature tracker description of the algorithm". In: *Intel Corporation* 5.1-10, p. 4.

Briane, Vincent, Charles Kervrann, and Myriam Vimond (2017). "A statistical analysis of particle trajectories in living cells". In: *arXiv preprint arXiv:1707.01838*.

Brox, Thomas and Jitendra Malik (2010). "Object segmentation by long term analysis of point trajectories". In: *Computer Vision–ECCV 2010*, pp. 282–295.

Brox, Thomas and Jitendra Malik (2011). "Large displacement optical flow: descriptor matching in variational motion estimation". In: *IEEE transactions on pattern analysis and machine intelligence* 33.3, pp. 500–513.

Brox, Thomas, Christoph Bregler, and Jitendra Malik (2009). "Large displacement optical flow". In: *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*. IEEE, pp. 41–48.

Bruna, Joan et al. (2013). "Spectral networks and locally connected networks on graphs". In: *arXiv preprint arXiv:1312.6203*.

Buades, Antoni, Bartomeu Coll, and J-M Morel (2005). "A non-local algorithm for image denoising". In: *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*. Vol. 2. IEEE, pp. 60–65.

Buske, Peter et al. (2011). "A comprehensive model of the spatio-temporal stem cell and tissue organisation in the intestinal crypt". In: *PLoS computational biology* 7.1, e1001045.

Buske, Peter et al. (2012). "On the biomechanics of stem cell niche formation in the gut–modelling growing organoids". In: *The FEBS journal* 279.18, pp. 3475–3487.

Butler, Daniel J et al. (2012). "A naturalistic open source movie for optical flow evaluation". In: *European Conference on Computer Vision*. Springer, pp. 611–625.

Caetano, Carlos Antônio et al. (2017). "Activity Recognition based on a Magnitude-Orientation Stream Network". In: *Graphics, Patterns and Images (SIBGRAPI), 2017 30th SIBGRAPI Conference on*. IEEE, pp. 47–54.

Cai, Anna Q, Kerry A Landman, and Barry D Hughes (2007). "Multi-scale modeling of a wound-healing cell migration assay." In: *Journal of theoretical biology* 245 (3), pp. 576–594.

Cai, Zhuowei et al. (2014). "Multi-view super vector for action recognition". In: *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pp. 596–603.

Cantrell, Michael A and Calvin J Kuo (2015). "Organoid modeling for cancer precision medicine". In: *Genome medicine* 7.1, p. 32.

Carabotti, Marilia et al. (2015). "The gut-brain axis: interactions between enteric microbiota, central and enteric nervous systems". In: *Annals of gastroenterology: quarterly publication of the Hellenic Society of Gastroenterology* 28.2, p. 203.

Carreira, Joao and Cristian Sminchisescu (2010). "Constrained parametric min-cuts for automatic object segmentation". In: *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*. IEEE, pp. 3241–3248.

Carreira, Joao and Andrew Zisserman (2017). "Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset". In: *arXiv preprint arXiv:1705.07750*.

Cavagna, Andrea et al. (2010). "Scale-free correlations in starling flocks". In: *Proceedings of the National Academy of Sciences* 107.26, pp. 11865–11870.

Cerbino, Roberto and Veronique Trappe (2008). "Differential dynamic microscopy: probing wave vector dependent dynamics with a microscope". In: *Physical review letters* 100.18, p. 188102.

Chang, Ming-Ching, Nils Krahnstoever, and Weina Ge (2011). "Probabilistic group-level motion analysis and scenario recognition". In: *Computer Vision (ICCV), 2011 IEEE International Conference on*. IEEE, pp. 747–754.

Chenouard, Nicolas et al. (2014). "Objective comparison of particle tracking methods". In: *Nature methods* 11.3, pp. 281–289.

Chollet, Francois (2017a). *The future of deep learning.* `https://blog.keras.io/the-future-of-deep-learning.html`. Blog.

— (2017b). *The limitations of deep learning.* `https://blog.keras.io/the-limitations-of-deep-learning.html`. Blog.

Clarke, G et al. (2013). "The microbiome-gut-brain axis during early life regulates the hippocampal serotonergic system in a sex-dependent manner". In: *Molecular psychiatry* 18.6, p. 666.

Comaniciu, Dorin and Peter Meer (2002). "Mean shift: A robust approach toward feature space analysis". In: *IEEE Transactions on pattern analysis and machine intelligence* 24.5, pp. 603–619.

Cronin, James et al. (2011). "Epidermal growth factor receptor (EGFR) is overexpressed in high-grade dysplasia and adenocarcinoma of the esophagus and may represent a biomarker of histological progression in Barrett's esophagus (BE)". In: *The American journal of gastroenterology* 106.1, pp. 46–56.

Dahmann, Christian, Andrew C Oates, and Michael Brand (2011). "Boundary formation and maintenance in tissue development". In: *Nature Reviews Genetics* 12.1 (1), pp. 43–55.

Dalal, Navneet and Bill Triggs (2005). "Histograms of oriented gradients for human detection". In: *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on.* Vol. 1. IEEE, pp. 886–893.

Dalal, Navneet, Bill Triggs, and Cordelia Schmid (2006). "Human detection using oriented histograms of flow and appearance". In: *European conference on computer vision.* Springer, pp. 428–441.

Das, Tamal et al. (2015). "A molecular mechanotransduction pathway regulates collective migration of epithelial cells". In: *Nature cell biology* 17.3 (3), pp. 276–287.

Defferrard, Michaël, Xavier Bresson, and Pierre Vandergheynst (2016). "Convolutional neural networks on graphs with fast localized spectral filtering". In: *Advances in Neural Information Processing Systems*, pp. 3844–3852.

Deforet, Maxime et al. (2012). "Automated velocity mapping of migrating cell populations (AVeMap)". In: *Nature methods* 9.11, pp. 1081–1083.

Dekkers, Johanna F et al. (2013). "A functional CFTR assay using primary cystic fibrosis intestinal organoids". In: *Nature medicine* 19.7, pp. 939–945.

Deng, Jia et al. (2009). "Imagenet: A large-scale hierarchical image database". In: *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on.* IEEE, pp. 248–255.

Dieterich, Peter et al. (2008). "Anomalous dynamics of cell migration". In: *Proceedings of the National Academy of Sciences* 105.2, pp. 459–463.

Dixon, MF et al. (2001). "Bile reflux gastritis and Barrett's oesophagus: further evidence of a role for duodenogastro-oesophageal reflux?" In: *Gut* 49.3, pp. 359–363.

Dosovitskiy, Alexey et al. (2015). "Flownet: Learning optical flow with convolutional networks". In: *Proceedings of the IEEE International Conference on Computer Vision*, pp. 2758–2766.

Dossa, Avafia Y et al. (2015). "Bile acids regulate intestinal cell proliferation by modulating EGFR and FXR signaling". In: *American Journal of Physiology-Gastrointestinal and Liver Physiology.*

Drost, Jarno et al. (2015). "Sequential cancer mutations in cultured human intestinal stem cells". In: *Nature* 521.7550, p. 43.

Duvenaud, David K et al. (2015). "Convolutional networks on graphs for learning molecular fingerprints". In: *Advances in neural information processing systems*, pp. 2224–2232.

Ellsworth, Rachel E et al. (2017). "Molecular heterogeneity in breast cancer: state of the science and implications for patient care". In: *Seminars in cell & developmental biology*. Vol. 64. Elsevier, pp. 65–72.

Endres, Ian and Derek Hoiem (2010). "Category independent object proposals". In: *Computer Vision–ECCV 2010*, pp. 575–588.

Evans, MR et al. (1995). "Asymmetric exclusion model with two species: spontaneous symmetry breaking". In: *Journal of statistical physics* 80.1, pp. 69–102.

Evans, Richard PT et al. (2016). "Evolving management of metaplasia and dysplasia in Barrett's epithelium". In: *World journal of gastroenterology* 22.47, p. 10316.

Everingham, Mark et al. (2010). "The pascal visual object classes (voc) challenge". In: *International journal of computer vision* 88.2, pp. 303–338.

Failmezger, Henrik, Holger Fröhlich, and Achim Tresch (2013). "Unsupervised automated high throughput phenotyping of RNAi time-lapse movies". In: *BMC bioinformatics* 14.1, p. 292.

Farnebäck, Gunnar (2003). "Two-frame motion estimation based on polynomial expansion". In: *Image analysis*, pp. 363–370.

Fatehullah, Aliya, Si Hui Tan, and Nick Barker (2016). "Organoids as an in vitro model of human development and disease". In: *Nature cell biology* 18.3, pp. 246–254.

Feichtenhofer, Christoph, Axel Pinz, and Andrew Zisserman (2016). "Convolutional two-stream network fusion for video action recognition". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1933–1941.

Felzenszwalb, Pedro F and Daniel P Huttenlocher (2004). "Efficient graph-based image segmentation". In: *International journal of computer vision* 59.2, pp. 167–181.

Fernandez, Romain et al. (2010). "Imaging plant growth in 4D: robust tissue reconstruction and lineaging at cell resolution". In: *Nature methods* 7.7, pp. 547–553.

Fey, Stephen J and Krzysztof Wrzesinski (2012). "Determination of drug toxicity using 3D spheroids constructed from an immortal human hepatocyte cell line". In: *Toxicological sciences* 127.2, pp. 403–411.

Finkbeiner, Stacy R et al. (2015). "Transcriptome-wide analysis reveals hallmarks of human intestine development and maturation in vitro and in vivo". In: *Stem cell reports* 4.6, pp. 1140–1155.

Fletcher, Alexander G et al. (2014). "Vertex models of epithelial morphogenesis." In: *Biophysical journal* 106 (11), pp. 2291–2304.

Fordham, Robert P et al. (2013). "Transplantation of expanded fetal intestinal progenitors contributes to colon regeneration after injury". In: *Cell stem cell* 13.6, pp. 734–744.

Fortun, Denis, Patrick Bouthemy, and Charles Kervrann (2015). "Optical flow modeling and computation: a survey". In: *Computer Vision and Image Understanding* 134, pp. 1–21.

Foster, Jane A and Karen-Anne McVey Neufeld (2013). "Gut–brain axis: how the microbiome influences anxiety and depression". In: *Trends in neurosciences* 36.5, pp. 305–312.

Fox, Emily B et al. (2007). "The sticky HDP-HMM: Bayesian nonparametric hidden Markov models with persistent states". In: *Arxiv preprint*.

Fragkiadaki, Katerina, Geng Zhang, and Jianbo Shi (2012). "Video segmentation by tracing discontinuities in a trajectory embedding". In: *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*. IEEE, pp. 1846–1853.

Fraley, Chris and Adrian E Raftery (1998). "How many clusters? Which clustering method? Answers via model-based cluster analysis". In: *The computer journal* 41.8, pp. 578–588.

Friedl, Peter (2004). "Prespecification and plasticity: shifting mechanisms of cell migration". In: *Current opinion in cell biology* 16.1, pp. 14–23.

Friedl, Peter and Bettina Weigelin (2008). "Interstitial leukocyte migration and immune function". In: *Nature immunology* 9.9, pp. 960–969.

Friedl, Peter et al. (2012). "Classifying collective cancer cell invasion". In: *Nature cell biology* 14.8, p. 777.

Gaddam, Srinivas et al. (2013). "Persistence of nondysplastic Barrett's esophagus identifies patients at lower risk for esophageal adenocarcinoma: results from a large multicenter cohort". In: *Gastroenterology* 145.3 (3), 548–553. e1.

Garcia-Garcia, Alberto et al. (2017). "A Review on Deep Learning Techniques Applied to Semantic Segmentation". In: *arXiv preprint arXiv:1704.06857*.

Garvey, Colleen M et al. (2016). "A high-content image-based method for quantitatively studying context-dependent cell population dynamics". In: *Scientific reports* 6, p. 29752.

Geiger, Andreas, Philip Lenz, and Raquel Urtasun (2012). "Are we ready for autonomous driving? the kitti vision benchmark suite". In: *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*. IEEE, pp. 3354–3361.

Giraud, Rémi et al. (2017). "SuperPatchMatch: an algorithm for robust correspondences using superpixel patches". In: *IEEE Transactions on Image Processing* 26.8, pp. 4068–4078.

Girshick, Ross (2015). "Fast r-cnn". In: *Proceedings of the IEEE international conference on computer vision*, pp. 1440–1448.

Girshick, Ross et al. (2014). "Rich feature hierarchies for accurate object detection and semantic segmentation". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 580–587.

Godinez, William J et al. (2011). "Tracking multiple particles in fluorescence microscopy images via probabilistic data association". In: *Biomedical Imaging: From Nano to Macro, 2011 IEEE International Symposium on*. IEEE, pp. 1925–1928.

Goodfellow, Ian et al. (2014a). "Generative adversarial nets". In: *Advances in neural information processing systems*, pp. 2672–2680.

Goodfellow, Ian J, Jonathon Shlens, and Christian Szegedy (2014b). "Explaining and harnessing adversarial examples". In: *arXiv preprint arXiv:1412.6572*.

Green, Melissa A, Clarence W Rowley, and Alexander J Smits (2010). "Using hyperbolic Lagrangian coherent structures to investigate vortices in bioinspired fluid flows". In: *Chaos: An Interdisciplinary Journal of Nonlinear Science* 20.1, p. 017510.

Green, Melville S (1954). "Markoff random processes and the statistical mechanics of time-dependent phenomena. II. Irreversible processes in fluids". In: *The Journal of Chemical Physics* 22.3, pp. 398–413.

Grégoire, Guillaume and Hugues Chaté (2004). "Onset of collective and cohesive motion". In: *Physical review letters* 92.2, p. 025702.

Guizar-Sicairos, Manuel, Samuel T Thurman, and James R Fienup (2008). "Efficient subpixel image registration algorithms". In: *Optics letters* 33.2, pp. 156–158.

Haller, George (2002). "Lagrangian coherent structures from approximate velocity data". In: *Physics of fluids* 14.6, pp. 1851–1861.

Hanahan, Douglas and Robert A Weinberg (2011). "Hallmarks of cancer: the next generation". In: *cell* 144.5, pp. 646–674.

Harada, Hideki et al. (2003). "Telomerase Induces Immortalization of Human Esophageal Keratinocytes Without p16INK4a Inactivation". In: *Molecular Cancer Research* 1.10, pp. 729–738.

Harel, Jonathan, Christof Koch, and Pietro Perona (2007). "Graph-based visual saliency". In: *Advances in neural information processing systems*, pp. 545–552.

Hariharan, Bharath et al. (2014). "Simultaneous detection and segmentation". In: *European Conference on Computer Vision*. Springer, pp. 297–312.

Hatzikirou, Haralambos and Andreas Deutsch (2008). "Cellular automata as microscopic models of cell migration in heterogeneous environments". In: *Current topics in developmental biology* 81, pp. 401–434.

He, Kaiming et al. (2016). "Deep residual learning for image recognition". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778.

He, Tao et al. (2017). "Cell tracking using deep neural networks with multi-task learning". In: *Image and Vision Computing* 60, pp. 142–153.

Hearst, Marti A. et al. (1998). "Support vector machines". In: *IEEE Intelligent Systems and their applications* 13.4, pp. 18–28.

Helbing, Dirk and Peter Molnar (1995). "Social force model for pedestrian dynamics". In: *Physical review E* 51.5, p. 4282.

Held, Michael et al. (2010). "CellCognition: time-resolved phenotype annotation in high-throughput live cell imaging". In: *Nature methods* 7.9, pp. 747–754.

Henaff, Mikael, Joan Bruna, and Yann LeCun (2015). "Deep convolutional networks on graph-structured data". In: *arXiv preprint arXiv:1506.05163*.

Henry, Katherine M et al. (2013). "PhagoSight: an open-source MATLAB® package for the analysis of fluorescent neutrophil and macrophage migration in a zebrafish model". In: *PloS one* 8.8, e72636.

Hilsenbeck, Oliver et al. (2016). "Software tools for single-cell tracking and quantification of cellular and molecular properties". In: *Nature biotechnology* 34.7, pp. 703–706.

Horn, Berthold KP and Brian G Schunck (1981). "Determining optical flow". In: *Artificial intelligence* 17.1-3, pp. 185–203.

Huang, Chang, Bo Wu, and Ramakant Nevatia (2008). "Robust object tracking by hierarchical association of detection responses". In: *European Conference on Computer Vision*. Springer, pp. 788–801.

Huang, Chang, Yuan Li, and Ramakant Nevatia (2013). "Multiple target tracking by learning-based hierarchical association of detection responses". In: *IEEE transactions on pattern analysis and machine intelligence* 35.4, pp. 898–910.

Huang, Gary B et al. (2007). *Labeled faces in the wild: A database for studying face recognition in unconstrained environments*. Tech. rep. Technical Report 07-49, University of Massachusetts, Amherst.

Huang, Ling et al. (2015). "Ductal pancreatic cancer modeling and drug screening using human pluripotent stem cell-and patient-derived tumor organoids". In: *Nature medicine* 21.11, pp. 1364–1371.

Huch, Meritxell and Bon-Kyoung Koo (2015). "Modeling mouse and human development using organoid cultures". In: *Development* 142.18, pp. 3113–3125.

Huszár, Ferenc (2016). *How powerful are Graph Convolutions? (review of Kipf & Welling, 2016)*. `http://www.inference.vc/how-powerful-are-graph-convolutions-review-of-kipf-welling-2016-2/`. Blog.

Ilg, Eddy et al. (2016). "Flownet 2.0: Evolution of optical flow estimation with deep networks". In: *arXiv preprint arXiv:1612.01925*.

Imamura, Yoshinori et al. (2015). "Comparison of 2D-and 3D-culture models as drug-testing platforms in breast cancer". In: *Oncology reports* 33.4, pp. 1837–1843.

Itti, Laurent and Pierre Baldi (2005). "A principled approach to detecting surprising events in video". In: *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*. Vol. 1. IEEE, pp. 631–637.

Itti, Laurent, Christof Koch, and Ernst Niebur (1998). "A model of saliency-based visual attention for rapid scene analysis". In: *IEEE Transactions on pattern analysis and machine intelligence* 20.11, pp. 1254–1259.

Jain, Ashesh et al. (2016a). "Structural-RNN: Deep learning on spatio-temporal graphs". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5308–5317.

Jain, Mihir, Hervé Jégou, and Patrick Bouthemy (2016b). "Improved motion description for action classification". In: *Frontiers in ICT* 2, p. 28.

Jaqaman, Khuloud et al. (2008). "Robust single-particle tracking in live-cell time-lapse sequences". In: *Nature methods* 5.8, pp. 695–702.

Jiang, Yu-Gang et al. (2017). "Exploiting feature and class relationships in video categorization with regularized deep neural networks". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Karpathy, Andrej et al. (2014). "Large-scale video classification with convolutional neural networks". In: *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pp. 1725–1732.

Kasten, Jens et al. (2010). "Lagrangian feature extraction of the cylinder wake". In: *Physics of Fluids (1994-present)* 22.9, p. 091108.

Kedrin, Dmitriy et al. (2008). "Intravital imaging of metastatic behavior through a mammary imaging window". In: *Nature methods* 5.12, pp. 1019–1021.

Kipf, Thomas (2016). *Graph Convolutional Networks*. `https://tkipf.github.io/graph-convolutional-networks/#fn2`. Blog.

Kipf, Thomas N and Max Welling (2016). "Semi-supervised classification with graph convolutional networks". In: *arXiv preprint arXiv:1609.02907*.

Kirkpatrick, James et al. (2017). "Overcoming catastrophic forgetting in neural networks". In: *Proceedings of the National Academy of Sciences*, p. 201611835.

Klaser, Alexander, Marcin Marszałek, and Cordelia Schmid (2008). "A spatio-temporal descriptor based on 3d-gradients". In: *BMVC 2008-19th British Machine Vision Conference*. British Machine Vision Association, pp. 275–1.

Klingemann, Mario (2017). *Rasterfairy*. `https://github.com/Quasimondo/RasterFairy`. GitHub Repository.

Knuth, Donald E (1997). *The art of computer programming, Volume 1 Fundamental Algorithms*. Addison-Wesley.

Koslover, Elena F, Caleb K Chan, and Julie A Theriot (2016). "Disentangling Random Motion and Flow in a Complex Medium". In: *Biophysical journal* 110.3, pp. 700–709.

Krizhevsky, Alex, Ilya Sutskever, and Geoffrey E Hinton (2012). "Imagenet classification with deep convolutional neural networks". In: *Advances in neural information processing systems*, pp. 1097–1105.

Kroeger, Till et al. (2016). "Fast optical flow using dense inverse search". In: *European Conference on Computer Vision*. Springer, pp. 471–488.

Ku, Tien-Chuan et al. (2007). "An automated tracking system to measure the dynamic properties of vesicles in living cells". In: *Microscopy research and technique* 70.2, pp. 119–134.

Kubo, Ryogo (1957). "Statistical-mechanical theory of irreversible processes. I. General theory and simple applications to magnetic and conduction problems". In: *Journal of the Physical Society of Japan* 12.6, pp. 570–586.

Kuehne, Hilde et al. (2013). "HMDB51: A large video database for human motion recognition". In: *High Performance Computing in Science and Engineering '12*. Springer, pp. 571–582.

Kuhn, Harold W (1955). "The Hungarian method for the assignment problem". In: *Naval Research Logistics (NRL)* 2.1-2, pp. 83–97.

Landsberg, Katharina P et al. (2009). "Increased cell bond tension governs cell sorting at the Drosophila anteroposterior compartment boundary". In: *Current Biology* 19.22, pp. 1950–1955.

Laptev, Ivan et al. (2008). "Learning realistic human actions from movies". In: *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*. IEEE, pp. 1–8.

Leal-Taix, Laura et al. (2017). "Tracking the trackers: an analysis of the state of the art in multiple object tracking". In: *arXiv preprint arXiv:1704.02781*.

LeCun, Yann et al. (1998). "Gradient-based learning applied to document recognition". In: *Proceedings of the IEEE* 86.11, pp. 2278–2324.

Lekien, Francois and Shane D Ross (2010). "The computation of finite-time Lyapunov exponents on unstructured meshes and for non-Euclidean manifolds". In: *Chaos: An Interdisciplinary Journal of Nonlinear Science* 20.1, p. 017505.

Li, Fuhai et al. (2010). "Multiple nuclei tracking using integer programming for quantitative cancer cell cycle analysis". In: *IEEE transactions on medical imaging* 29.1, pp. 96–105.

Li, Li et al. (2012). "Live imaging reveals differing roles of macrophages and neutrophils during zebrafish tail fin regeneration". In: *Journal of Biological Chemistry* 287.30, pp. 25353–25360.

Li, Liang, Edward C Cox, and Henrik Flyvbjerg (2011). "'Dicty dynamics': Dictyostelium motility as persistent random motion". In: *Physical biology* 8.4, p. 046006.

Li, Xiaoxiao et al. (2017). "Video Object Segmentation with Re-identification". In: *arXiv preprint arXiv:1708.00197*.

Li, Xuelong, Mulin Chen, and Qi Wang (2016a). "Measuring collectiveness via refined topological similarity". In: *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)* 12.2, p. 34.

Li, Yi et al. (2016b). "Fully convolutional instance-aware semantic segmentation". In: *arXiv preprint arXiv:1611.07709*.

Li, Yujia et al. (2015). "Gated graph sequence neural networks". In: *arXiv preprint arXiv:1511.05493*.

Liang, Chun-Chi, Ann Y Park, and Jun-Lin Guan (2007). "In vitro scratch assay: a convenient and inexpensive method for analysis of cell migration in vitro". In: *Nature protocols* 2.2, pp. 329–333.

Lin, Tsung-Yi et al. (2014). "Microsoft coco: Common objects in context". In: *European conference on computer vision*. Springer, pp. 740–755.

Liu, Ce, Jenny Yuen, and Antonio Torralba (2011). "Sift flow: Dense correspondence across scenes and its applications". In: *IEEE transactions on pattern analysis and machine intelligence* 33.5, pp. 978–994.

Lou, Xinghua and Fred A Hamprecht (2011). "Structured learning for cell tracking". In: *Advances in neural information processing systems*, pp. 1296–1304.

Lowe, David G (1999). "Object recognition from local scale-invariant features". In: *Computer vision, 1999. The proceedings of the seventh IEEE international conference on*. Vol. 2. Ieee, pp. 1150–1157.

Lucas, Bruce D and Takeo Kanade (1981). "An iterative image registration technique with an application to stereo vision". In:

Luo, Wenhan et al. (2014). "Multiple object tracking: A literature review". In: *arXiv preprint arXiv:1409.7618*.

Magnusson, Klas EG et al. (2015). "Global linking of cell tracks using the Viterbi algorithm". In: *IEEE transactions on medical imaging* 34.4, pp. 911–929.

Maire, Michael et al. (2008). "Using contours to detect and localize junctions in natural images". In: *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*. IEEE, pp. 1–8.

Major, Robert J and Kenneth D Irvine (2005). "Influence of Notch on dorsoventral compartmentalization and actin organization in the Drosophila wing". In: *Development* 132.17, pp. 3823–3833.

— (2006). "Localization and requirement for Myosin II at the dorsal-ventral compartment boundary of the Drosophila wing". In: *Developmental dynamics* 235.11, pp. 3051–3058.

Mallet, Daniel G and Lisette G De Pillis (2006). "A cellular automata model of tumor-immune system interactions". In: *Journal of Theoretical Biology* 239.3, pp. 334–350.

Manessi, Franco, Alessandro Rozza, and Mario Manzo (2017). "Dynamic Graph Convolutional Networks". In: *arXiv preprint arXiv:1704.06199*.

Marel, Anna-Kristina et al. (2014a). "Alignment of cell division axes in directed epithelial cell migration". In: *New Journal of Physics* 16.11, p. 115005.

Marel, Anna-Kristina et al. (2014b). "Flow and diffusion in channel-guided cell migration". In: *Biophysical journal* 107.5, pp. 1054–1064.

Markham, Deborah C, Matthew J Simpson, and Ruth E Baker (2015). "Choosing an appropriate modelling framework for analysing multispecies co-culture cell biology experiments". In: *Bulletin of mathematical biology* 77.4 (4), pp. 713–734.

Martin, David R, Charless C Fowlkes, and Jitendra Malik (2004). "Learning to detect natural image boundaries using local brightness, color, and texture cues". In: *IEEE transactions on pattern analysis and machine intelligence* 26.5, pp. 530–549.

Maška, Martin et al. (2014). "A benchmark for comparison of cell tracking algorithms". In: *Bioinformatics* 30.11, pp. 1609–1617.

Masuzzo, Paola et al. (2016). "Taking aim at moving targets in computational cell migration". In: *Trends in cell biology* 26.2 (2), pp. 88–110.

Mayer, Nikolaus et al. (2016). "A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4040–4048.

Méhes, Előd and Tamás Vicsek (2014). "Collective motion of cells: from experiments to models". In: *Integrative biology* 6.9, pp. 831–854.

Mehran, Ramin, Alexis Oyama, and Mubarak Shah (2009). "Abnormal crowd behavior detection using social force model". In: *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*. IEEE, pp. 935–942.

Meijering, Erik et al. (2009). "Tracking in cell and developmental biology". In: *Seminars in cell & developmental biology*. Vol. 20. 8. Elsevier, pp. 894–902.

Meijering, Erik, Oleh Dzyubachyk, and Ihor Smal (2012). "9 Methods for Cell and Particle Tracking". In: *Methods in enzymology* 504.9, pp. 183–200.

Menze, Moritz and Andreas Geiger (2015). "Object scene flow for autonomous vehicles". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3061–3070.

Merlo, Lauren MF et al. (2011). "An in vitro co-culture model of esophageal cells identifies ascorbic acid as a modulator of cell competition". In: *BMC cancer* 11.1, p. 461.

Metz, Luke et al. (2016). "Unrolled generative adversarial networks". In: *arXiv preprint arXiv:1611.02163*.

Milde, Florian et al. (2012). "Cell Image Velocimetry (CIV): boosting the automated quantification of cell migration in wound healing assays". In: *Integrative Biology* 4.11, pp. 1437–1447.

Mills, Richard J et al. (2017). "Functional screening in human cardiac organoids reveals a metabolic mechanism for cardiomyocyte cell cycle arrest". In: *Proceedings of the National Academy of Sciences*, p. 201707316.

Monier, Bruno et al. (2010). "An actomyosin-based barrier inhibits cell mixing at compartmental boundaries in Drosophila embryos". In: *Nature cell biology* 12.1, pp. 60–65.

Montell, Denise J (2008). "Morphogenetic cell movements: diversity from modular mechanical properties". In: *Science* 322.5907, pp. 1502–1505.

Nagy, Máté et al. (2010). "Hierarchical group dynamics in pigeon flocks". In: *arXiv preprint arXiv:1010.5394*.

Netzer, Yuval et al. (2011). "Reading digits in natural images with unsupervised feature learning". In: *NIPS workshop on deep learning and unsupervised feature learning*. Vol. 2011. 2, p. 5.

Neumann, Beate et al. (2006). "High-throughput RNAi screening by time-lapse imaging of live human cells". In: *Nature methods* 3.5, p. 385.

Nguyen, Anh, Jason Yosinski, and Jeff Clune (2015). "Deep neural networks are easily fooled: High confidence predictions for unrecognizable images". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 427–436.

Odena, Augustus, Christopher Olah, and Jonathon Shlens (2016). "Conditional image synthesis with auxiliary classifier gans". In: *arXiv preprint arXiv:1610.09585*.

Padfield, Dirk, Jens Rittscher, and Badrinath Roysam (2011). "Coupled minimum-cost flow cell tracking for high-throughput quantitative analysis". In: *Medical image analysis* 15.4, pp. 650–668.

Park, Jin-Ah et al. (2015). "Unjamming and cell shape in the asthmatic airway epithelium". In: *Nature materials* 14.10, pp. 1040–1048.

Patsialou, Antonia et al. (2013). "Intravital multiphoton imaging reveals multicellular streaming as a crucial component of in vivo cell migration in human breast tumors". In: *Intravital* 2.2, e25294.

Pauli, Chantal et al. (2017). "Personalized in vitro and in vivo cancer models to guide precision medicine". In: *Cancer Discovery* 7.5, pp. 462–477.

Peng, Jiangtao et al. (2010). "Asymmetric least squares for multiple spectra baseline correction". In: *Analytica chimica acta* 683.1, pp. 63–68.

Peng, Xiaojiang et al. (2016). "Bag of visual words and fusion methods for action recognition: Comprehensive study and good practice". In: *Computer Vision and Image Understanding* 150, pp. 109–125.

Perc, Matjaž (2005). "The dynamics of human gait". In: *European journal of physics* 26.3, p. 525.

Pertuz, Said, Domenec Puig, and Miguel Angel Garcia (2013). "Analysis of focus measure operators for shape-from-focus". In: *Pattern Recognition* 46.5, pp. 1415–1432.

Petitjean, Laurence et al. (2010). "Velocity fields in a collectively migrating epithelium". In: *Biophysical journal* 98.9, pp. 1790–1800.

Pinheiro, Pedro O, Ronan Collobert, and Piotr Dollár (2015). "Learning to segment object candidates". In: *Advances in Neural Information Processing Systems*, pp. 1990–1998.

Pinheiro, Pedro O et al. (2016). "Learning to refine object segments". In: *European Conference on Computer Vision*. Springer, pp. 75–91.

Pittet, Mikael J and Ralph Weissleder (2011). "Intravital imaging". In: *Cell* 147.5, pp. 983–991.

Podewitz, Nils et al. (2016). "Interface dynamics of competing tissues". In: *New Journal of Physics* 18.8, p. 083020.

Pont-Tuset, Jordi et al. (2017). "The 2017 davis challenge on video object segmentation". In: *arXiv preprint arXiv:1704.00675*.

Progatzky, Fränze, Margaret J Dallman, and Cristina Lo Celso (2013). "From seeing to believing: labelling strategies for in vivo cell-tracking experiments". In: *Interface focus* 3.3, p. 20130001.

Radford, Alec, Luke Metz, and Soumith Chintala (2015). "Unsupervised representation learning with deep convolutional generative adversarial networks". In: *arXiv preprint arXiv:1511.06434*.

Redmon, Joseph et al. (2016). "You only look once: Unified, real-time object detection". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 779–788.

Ren, Shaoqing et al. (2015). "Faster R-CNN: Towards real-time object detection with region proposal networks". In: *Advances in neural information processing systems*, pp. 91–99.

Research, Google (2016). *Zero-Shot Translation with Google's Multilingual Neural Machine Translation System*. `https://research.googleblog.com/2016/11/zero-shot-translation-with-googles.html`. Blog.

Revaud, Jerome et al. (2015). "Epicflow: Edge-preserving interpolation of correspondences for optical flow". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1164–1172.

Ridley, Anne J et al. (2003). "Cell migration: integrating signals from front to back". In: *Science* 302.5651, pp. 1704–1709.

Riedl, Angelika et al. (2017). "Comparison of cancer cells in 2D vs 3D culture reveals differences in AKT–mTOR–S6K signaling and drug responses". In: *J Cell Sci* 130.1, pp. 203–218.

Robertson, Anne L et al. (2014). "A zebrafish compound screen reveals modulation of neutrophil reverse migration as an anti-inflammatory mechanism". In: *Science translational medicine* 6.225, 225ra29–225ra29.

Robinson, Mark D, Davis J McCarthy, and Gordon K Smyth (2010). "edgeR: a Bioconductor package for differential expression analysis of digital gene expression data". In: *Bioinformatics* 26.1, pp. 139–140.

Robinson, Sean et al. (2015). "Segmentation of Image Data from Complex Organotypic 3D Models of Cancer Tissues with Markov Random Fields". In: *PloS one* 10.12, e0143798.

Ronneberger, Olaf, Philipp Fischer, and Thomas Brox (2015a). "U-net: Convolutional networks for biomedical image segmentation". In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, pp. 234–241.

— (2015b). "U-net: Convolutional networks for biomedical image segmentation". In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, pp. 234–241.

Rosenstein, Michael T, James J Collins, and Carlo J De Luca (1993). "A practical method for calculating largest Lyapunov exponents from small data sets". In: *Physica D: Nonlinear Phenomena* 65.1-2, pp. 117–134.

Rousseeuw, Peter J (1987). "Silhouettes: a graphical aid to the interpretation and validation of cluster analysis". In: *Journal of computational and applied mathematics* 20, pp. 53–65.

Ruocco, Giancarlo and Francesco Sette (1999). "The high-frequency dynamics of liquid water". In: *Journal of Physics: Condensed Matter* 11.24, R259.

Sato, Toshiro et al. (2009). "Single Lgr5 stem cells build crypt-villus structures in vitro without a mesenchymal niche". In: *Nature* 459.7244, p. 262.

Scharstein, Daniel and Richard Szeliski (2002). "A taxonomy and evaluation of dense two-frame stereo correspondence algorithms". In: *International journal of computer vision* 47.1-3, pp. 7–42.

Schiegg, Martin et al. (2015). "Graphical model for joint segmentation and tracking of multiple dividing cells". In: *Bioinformatics* 31.6, pp. 948–956.

Schindelin, Johannes et al. (2012). "Fiji: an open-source platform for biological-image analysis". In: *Nature methods* 9.7, pp. 676–682.

Schütz, G and E Domany (1993). "Phase transitions in an exactly soluble one-dimensional exclusion process". In: *Journal of statistical physics* 72.1, pp. 277–296.

Schütz, Gunter M (1997). "Exact solution of the master equation for the asymmetric exclusion process". In: *Journal of statistical physics* 88.1, pp. 427–445.

Scott, David W (2015). *Multivariate density estimation: theory, practice, and visualization*. John Wiley & Sons.

Scovanner, Paul and Marshall F Tappen (2009). "Learning pedestrian dynamics from the real world". In: *Computer Vision, 2009 IEEE 12th International Conference on*. IEEE, pp. 381–388.

Secrier, Maria et al. (2016). "Mutational signatures in esophageal adenocarcinoma define etiologically distinct subgroups with therapeutic relevance". In: *Nature Genetics* 48.10, pp. 1131–1141.

Shadden, Shawn C, Francois Lekien, and Jerrold E Marsden (2005). "Definition and properties of Lagrangian coherent structures from finite-time Lyapunov exponents in two-dimensional aperiodic flows". In: *Physica D: Nonlinear Phenomena* 212.3, pp. 271–304.

Shadden, Shawn C, John O Dabiri, and Jerrold E Marsden (2006). "Lagrangian analysis of fluid transport in empirical vortex ring flows". In: *Physics of Fluids* 18.4, p. 047105.

Sharif Razavian, Ali et al. (2014). "CNN features off-the-shelf: an astounding baseline for recognition". In: *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pp. 806–813.

Shi, Feng, Robert Laganiere, and Emil Petriu (2015). "Gradient boundary histograms for action recognition". In: *Applications of Computer Vision (WACV), 2015 IEEE Winter Conference on*. IEEE, pp. 1107–1114.

Shi, Jianbo et al. (1994). "Good features to track". In: *Computer Vision and Pattern Recognition, 1994. Proceedings CVPR'94., 1994 IEEE Computer Society Conference on*. IEEE, pp. 593–600.

Shin, Hanul et al. (2017). "Continual Learning with Deep Generative Replay". In: *arXiv preprint arXiv:1705.08690*.

Shishika, Daigo et al. (2014). "Male motion coordination in anopheline mating swarms". In: *Scientific reports* 4.

Simonyan, Karen and Andrew Zisserman (2014a). "Two-stream convolutional networks for action recognition in videos". In: *Advances in neural information processing systems*, pp. 568–576.

— (2014b). "Very deep convolutional networks for large-scale image recognition". In: *arXiv preprint arXiv:1409.1556*.

Simpson, Matthew J, Kerry A Landman, and Barry D Hughes (2009). "Multi-species simple exclusion processes". In: *Physica A: Statistical Mechanics and its Applications* 388.4, pp. 399–406.

Sinha, Sunil K (2001). "Theory of inelastic X-ray scattering from condensed matter". In: *Journal of Physics: Condensed Matter* 13.34, p. 7511.

Smeulders, Arnold WM et al. (2014). "Visual tracking: An experimental survey". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 36.7, pp. 1442–1468.

Soomro, Khurram, Amir Roshan Zamir, and Mubarak Shah (2012). "UCF101: A Dataset of 101 Human Actions Classes From Videos in The Wild". In: *CoRR* abs/1212.0402. URL: `http://arxiv.org/abs/1212.0402`.

Souza, Rhonda F (2010). *The role of acid and bile reflux in oesophagitis and Barrett's metaplasia*. Generic.

Srivastava, Nitish, Elman Mansimov, and Ruslan Salakhudinov (2015). "Unsupervised learning of video representations using lstms". In: *International Conference on Machine Learning*, pp. 843–852.

Sun, Chen et al. (2017). "Revisiting Unreasonable Effectiveness of Data in Deep Learning Era". In: *arXiv preprint arXiv:1707.02968*.

Sveen, Johan Kristian (2004). "An introduction to MatPIV v. 1.6. 1". In: *Preprint series. Mechanics and Applied Mathematics http://urn. nb. no/URN: NBN: no-23418*.

Szabo, Balint et al. (2006a). "Phase transition in the collective migration of tissue cells: experiment and model". In: *Physical Review E* 74.6, p. 061908.

Szabo, Balint et al. (2006b). "Phase transition in the collective migration of tissue cells: experiment and model". In: *Physical Review E* 74.6, p. 061908.

Szegedy, Christian et al. (2015). "Going deeper with convolutions". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1–9.

Takens, Floris et al. (1981). "Detecting strange attractors in turbulence". In: *Lecture notes in mathematics* 898.1, pp. 366–381.

Tambe, Dhananjay T et al. (2011). "Collective cell guidance by cooperative intercellular forces". In: *Nature materials* 10.6, p. 469.

Tan, Chin Wee, Yumiko Hirokawa, and Antony W Burgess (2015). "Analysis of Wnt signalling dynamics during colon crypt development in 3D culture". In: *Scientific reports* 5, p. 11036.

Tinevez, Jean-Yves et al. (2017). "TrackMate: An open and extensible platform for single-particle tracking". In: *Methods* 115, pp. 80–90.

Tomasi, Carlo and Takeo Kanade (1991). "Detection and tracking of point features". In:

Tomer, Raju et al. (2012). "Quantitative high-speed imaging of entire developing embryos with simultaneous multiview light-sheet microscopy". In: *Nature methods* 9.7, pp. 755–763.

Torralba, Antonio and Alexei A. Efros (2011). "Unbiased look at dataset bias". In: *The 24th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2011, Colorado Springs, CO, USA, 20-25 June 2011*. IEEE Computer Society, pp. 1521–1528.

Tran-Dinh, Quoc (2015). "Construction and Iteration-Complexity of Primal Sequences in Alternating Minimization Algorithms". In: arXiv: `1511.03305v1 [math.OC]`.

Trepat, Xavier et al. (2009). "Physical forces during collective cell migration". In: *Nature physics* 5.6, pp. 426–430.

Trichas, Georgios et al. (2012). "Multi-cellular rosettes in the mouse visceral endoderm facilitate the ordered migration of anterior visceral endoderm cells". In: *PLoS biology* 10.2, e1001256.

Uijlings, Jasper RR et al. (2013). "Selective search for object recognition". In: *International journal of computer vision* 104.2, pp. 154–171.

Uras, Sergio et al. (1988). "A computational approach to motion perception". In: *Biological Cybernetics* 60.2, pp. 79–87.

Van de Wetering, Marc et al. (2015). "Prospective derivation of a living organoid biobank of colorectal cancer patients." In: *Cell* 161.4 (4), pp. 933–945.

Van den Oord, Aaron et al. (2016). "Conditional image generation with pixelcnn decoders". In: *Advances in Neural Information Processing Systems*, pp. 4790–4798.

Van Valen, David A et al. (2016). "Deep learning automates the quantitative analysis of individual cells in live-cell imaging experiments". In: *PLoS computational biology* 12.11, e1005177.

Vaughan, RB and JP Trinkaus (1966). "Movements of epithelial cell sheets in vitro". In: *Journal of cell science* 1.4, pp. 407–413.

Vedula, Sri Ram Krishna et al. (2012). "Emerging modes of collective cell migration induced by geometrical constraints". In: *Proceedings of the National Academy of Sciences* 109.32, pp. 12974–12979.

Vicsek, Tamás and Anna Zafeiris (2012). "Collective motion". In: *Physics Reports* 517.3, pp. 71–140.

Vicsek, Tamás et al. (1995). "Novel type of phase transition in a system of self-driven particles". In: *Physical review letters* 75.6, p. 1226.

Vijayanarasimhan, Sudheendra et al. (2017). "Sfm-net: Learning of structure and motion from video". In: *arXiv preprint arXiv:1704.07804*.

Walsh, Alex J et al. (2016). "Drug response in organoids generated from frozen primary tumor tissues". In: *Scientific reports* 6.

Wang, Heng and Cordelia Schmid (2013). "Action recognition with improved trajectories". In: *Proceedings of the IEEE international conference on computer vision*, pp. 3551–3558.

Wang, Heng et al. (2011a). "Action recognition by dense trajectories". In: *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*. IEEE, pp. 3169–3176.

— (2013). "Dense trajectories and motion boundary descriptors for action recognition". In: *International journal of computer vision* 103.1, pp. 60–79.

Wang, Limin et al. (2016). "Temporal segment networks: Towards good practices for deep action recognition". In: *European Conference on Computer Vision*. Springer, pp. 20–36.

Wang, Naiyan and Dit-Yan Yeung (2013). "Learning a deep compact image representation for visual tracking". In: *Advances in neural information processing systems*, pp. 809–817.

Wang, Shu et al. (2011b). "Superpixel tracking". In: *Computer Vision (ICCV), 2011 IEEE International Conference on*. IEEE, pp. 1323–1330.

Weinzaepfel, Philippe et al. (2013). "DeepFlow: Large displacement optical flow with deep matching". In: *Proceedings of the IEEE International Conference on Computer Vision*, pp. 1385–1392.

Wu, Shandong, Brian E Moore, and Mubarak Shah (2010). "Chaotic invariants of lagrangian particle trajectories for anomaly detection in crowded scenes". In: *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*. IEEE, pp. 2054–2060.

Wu, Shandong, Omar Oreifej, and Mubarak Shah (2011). "Action recognition in videos acquired by a moving camera using motion decomposition of lagrangian particle trajectories". In: *Computer Vision (ICCV), 2011 IEEE International Conference on*. IEEE, pp. 1419–1426.

Wu, Yonghui et al. (2016). "Google's neural machine translation system: Bridging the gap between human and machine translation". In: *arXiv preprint arXiv:1609.08144*.

Xie, Weidi, J Alison Noble, and Andrew Zisserman (2016). "Microscopy cell counting and detection with fully convolutional regression networks". In: *Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization*, pp. 1–10.

Xu, Jia, René Ranftl, and Vladlen Koltun (2017). "Accurate optical flow via direct cost volume processing". In: *arXiv preprint arXiv:1704.07325*.

Yang, Fan, Huchuan Lu, and Ming-Hsuan Yang (2014). "Robust superpixel tracking". In: *IEEE Transactions on Image Processing* 23.4, pp. 1639–1651.

Yi, Song et al. (2017). "Functional variomics and network perturbation: connecting genotype to phenotype in cancer". In: *Nature Reviews Genetics* 18.7, pp. 395–410.

Yin, Xiaolei et al. (2014). "Niche-independent high-purity cultures of Lgr5+ intestinal stem cells and their progeny". In: *Nature methods* 11.1, pp. 106–112.

Yin, Xiaolei et al. (2016). "Engineering stem cell organoids". In: *Cell stem cell* 18.1, pp. 25–38.

Yosinski, Jason et al. (2014). "How transferable are features in deep neural networks?" In: *Advances in neural information processing systems*, pp. 3320–3328.

Yu, Bing, Haoteng Yin, and Zhanxing Zhu (2017). "Spatio-temporal Graph Convolutional Neural Network: A Deep Learning Framework for Traffic Forecasting". In: *arXiv preprint arXiv:1709.04875*.

Zagoruyko, Sergey et al. (2016). "A multipath network for object detection". In: *arXiv preprint arXiv:1604.02135*.

Zaritsky, Assaf et al. (2012a). "Emergence of HGF/SF-induced coordinated cellular motility". In: *PLoS One* 7.9, e44671.

— (2012b). "Emergence of HGF/SF-induced coordinated cellular motility". In: *PLoS One* 7.9, e44671.

Zhang, Yimeng et al. (2012). "Group context learning for event recognition". In: *Applications of Computer Vision (WACV), 2012 IEEE Workshop on*. IEEE, pp. 249–255.

Zhang, Zhang, Kaiqi Huang, and Tieniu Tan (2006). "Comparison of similarity measures for trajectory clustering in outdoor surveillance scenes". In: *Pattern Recognition, 2006. ICPR 2006. 18th International Conference on*. Vol. 3. IEEE, pp. 1135–1138.

Zheng, Shuai et al. (2015). "Conditional random fields as recurrent neural networks". In: *Proceedings of the IEEE International Conference on Computer Vision*, pp. 1529–1537.

Zhou, Bolei, Xiaoou Tang, and Xiaogang Wang (2013). "Measuring crowd collectiveness". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3049–3056.

Zhou, Denny et al. (2004). "Ranking on data manifolds". In: *Advances in neural information processing systems*, pp. 169–176.