Image Analysis and Platform Development for Automated Phenotyping in Cytomics

K. Yan

严 宽

Colophon

The studies described in this thesis were performed at the Leiden Institute of Advanced Computer Science in the Imaging & BioInformatics group, Leiden University, Leiden, The Netherlands.

The research was finanically supported by the BioRange Project of the Netherlands Bioinformatics Centre (NBIC).

Printed by Proefschriftmaken.nl || Uitgeverij BOXPress

Published by Uitgeverij BOXPress, 's-Hertogenbosch

ISBN 978-90-8891-762-2

Image Analysis and Platform Development for Automated Phenotyping in Cytomics

Proefschrift

ter verkrijging van

de graad van Doctor aan de Universiteit Leiden,

op gezag van Rector Magnificus Prof. Mr. C.J.J.M. Stolker,

volgens besluit van het College voor Promoties

te verdedigen op woensdag 27 November 2013

klokke 10:00 uur

door

Kuan Yan

geboren te Shanghai, China, in 1982

Promotiecommissie

Promotores:
Prof. Dr. J. N. Kok
Prof. Dr. B. Van de Water

Co-promotor:
Dr. Ir. F.J. Verbeek

Overige Leden:
Prof. Dr. X. Liu
Prof. Dr. H. Tanke
Prof. Dr. P. Lucas

The studies described in this thesis were performed at the Leiden Institute of Advanced Computer Science in the Imaging & BioInformatics group, Leiden University, Leiden, The Netherlands.

The research was finanically supported by the BioRange Project of the Netherlands Bioinformatics Centre (NBIC).

Table of Contents

Chapter 1

Introduction to High-throughput Cytomics

1.1. Bioinformatics and Cell Biology

Bioinformatics is the research field that attempts to extract comprehensive information from large quantities of biological data via an integration of methods from computer science, mathematics, and biology [1]. It is an interdisciplinary study of automation for the retrieval, storage and analysis of biological data. The term bioinformatics, so it is posed, was first in 1970 by two Dutch complex system researchers, Ben Hesper and Paulien Hogeweg [2]. Since then, bioinformatics has been the key to functional genetics and system biology experiment in order to understand mechanisms behind cell behavior and cell system. In current cell biology research, bioinformatics plays a crucial role in the correlation modeling between cell **genotype** and **phenotype**.

In cell biology, the term **genotype** refers to the unique genetic hereditary and expression pattern while the term **phenotype** refers to observable cell properties including morphology and migration behavior. The correlation modeling of **genotype-to-phenotype** is a study of the correlation/causality between characteristic patterns of gene expression and recognizable cell properties. The genotype-to-phenotype correlation is the foundation of a systematic understanding of molecular architecture and control mechanisms behind different cell behavior. This correlation is of particular importance to cell behavior related diseases. However, due to the amount of information to be analyzed, the study of genotype-to-phenotype correlation must be done in an automated fashion.

To that end, modern developments in quantitative microscopy and laboratory robotics have provided the necessary hardware to the automated study of genotype-to-phenotype correlation. These studies are often referred as the –omics studies[3]. Depending on the study subject and biology of interest, omics studies employed by molecular genetics and cell biology branched out to a number of fields including genomics, **transcriptomics**, proteomics, metabolomics and **cytomics**.

Transcriptomics is the study dedicated to the quantitative analysis of RNA expression. In a living cell, RNA is produced by transcripting DNA strains. The expression pattern of RNA is considered an indication of protein production related to cell behavior. The study of the correlation between RNA expression and cell behavior is an essential part of genotype-to-phenotype study. Much of it is accomplished via the application of an high-throughput screen technology known as the **microarray**. The microarray technology is a method originally derived from Southern blotting [4] developed by Sir Edwin Mellor Southern in 1992. Essentially, microarray technology relies on base-pairing and hybridization of RNA strains to quantify the expression level of targeted RNA strains.

Transcriptomics focuses on the quantification of the genotype, whereas **cytomics** is dedicated to the quantitative analysis of the cell phenotype using high-throughput screen technology; this is known as the high-throughput/high-content cell screen or simply **HT/HC screen**. The HT/HC screen is achieved via the introduction of automated microscopy. Modern HT/HC screen analysis produces data on terabyte scale. Such a data volume poses great difficulty in manual analysis in early cytomics studies. To that end, computer science and information theory including image analysis, pattern recognition and machine learning have been included

to provide automated solution for image data processing. Due to the unique characteristics of biological image data, image analysis solutions need to be tailored. Therefore, we systematically investigate the limitations of generic image analysis solutions in HT/HC screen experiments; and based on the limitation studies, this thesis will focus on the design of dedicated image analysis solutions for HT/HC screen studies.

1.2. Image Analysis in High-throughput/High-Content Screen

At the current stage, cytomic experiments frequently employs either **flow cytometry** or **high-content screen** techniques to capture cell phenotypes.

Flow cytometry (cf. Figure 1-1) is a laser based technology employed in cell counting, sorting [5][6], and biomarker detection by suspending cells in a fluid stream and passing them through an electronic detection apparatus.

The **high-throughput/high-content screen** (HT/HC screen) [7] is a microcopy based method aiming for the visualization and capture of cell phenotypes over a large range of experimental settings (cf. Figure 1-2 & Figure 1-4). With different microscopy techniques, the HT/HC screen can serve different biological studies.

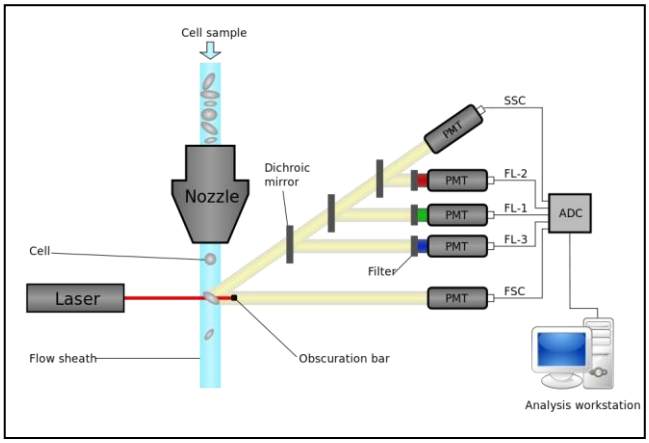


Figure 1-1 the workflow of a multi-color flow cytometry

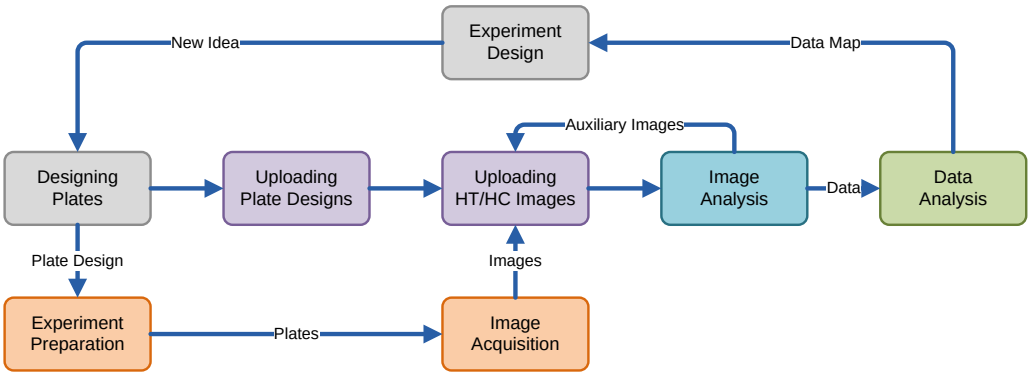


Figure 1-2 the general workflow of HT/HC screen study

In cytomics studies, the flow cytometry technology is frequently used to sort and quantify different types of cells [5][6][8] due to its fast analysis and high sorting accuracy. However, the flow cytometry does not provide the possible to study live cell behavior. Compared to flow cytometry, HT/HC screen using automated microscopy can capture both cell morphology and cell migration, i.e., size, polarization and migration speed. The employment of an HT/HC screen allows the extraction of phenotypical quantification; combined with microarray technology it

allows further correlation modeling of genotype-to-phenotype can be accomplished[9][10][11][12][13][14].

An HT/HC screen study consists of a sequence of five steps[9][14]: (1) **experiment design**, (2) **experiment preparation**, (3) **image acquisition**, (4) **image analysis**, and (5) **data analysis**. In each step, the raw data are further transformed into comprehensive data representation to support validation of the initial hypothesis.

1. Experiment Design

The experiment design is the first step in a HT/HC screen. It is the step to draw the blueprint for the HT/HC screen study to serve the research question. During the experiment design, the researchers must answer two essential questions: (1) what information must be captured and (2) how to capture the information. It is very difficult to be precise on how these two questions should be answered. Instead, we will give two examples of experiment design to explain the procedure.

Example 1: Growth factor regulated cancer metastasis [10][15] (cf. Figure 1-6a & Ch. 4)

In the experiment design of this HT/HC study, the researchers must first select a siRNA library which is potentially related to cell migration. Second, they must choose a cancer cell line demonstrating strong phenotypical plasticity, thus allowing to capture more observable differences in cell phenotype. Third, they must design a layout (cf. Figure 1-3) to fit the treated cells into a culture plate (cf. Figure 1-5). At the same time it is decided which microscopy technique is most suitable to capture the cell phenotype.

Example 2: Study of matrix adhesion dynamics [16] (cf. Figure 1-6b & Ch. 5)

In the experiment design of this HT/HC study, the researchers must first decide what combination of microscopy techniques can be used to capture matrix adhesion (subcellular structure) and cell migration (cf. Figure 1-4). Finally, they must decide how to quantify the relationship between the dynamics of matrix adhesions and cell migration.

From these two examples it is clear that for different biological questions, the experiment design step may consist of different tasks. As a result, there are very few limitations to the step of experiment design which makes it vulnerable to user generated error. In Chapter 6, we will demonstrate a flexible computer-aided experiment design interface for HT/HC screen studies.

DMSO	HGF	EGF+HGF	FGF+TGFbeta	DMSO	HGF	EGF+HGF	FGF+TGFbeta
DMSO	HGF	EGF+HGF	FGF+TGFbeta	DMSO	HGF	EGF+HGF	FGF+TGFbeta
EGF	TGFbeta	EGF+TGFbeta	HGF+TGFbeta	EGF	TGFbeta	EGF+TGFbeta	HGF+TGFbeta
EGF	TGFbeta	EGF+TGFbeta	HGF+TGFbeta	EGF	TGFbeta	EGF+TGFbeta	HGF+TGFbeta
FGF	EGF+FGF	FGF+HGF	all	FGF	EGF+FGF	FGF+HGF	all
FGF	EGF+FGF	FGF+HGF	all	FGF	EGF+FGF	FGF+HGF	all
pIGFP				GB1			

Figure 1-3 plate layout of a 8x12 culture plate used in RCM3 growth factor regulation (cf. chapter 4) in Excel spreadsheet

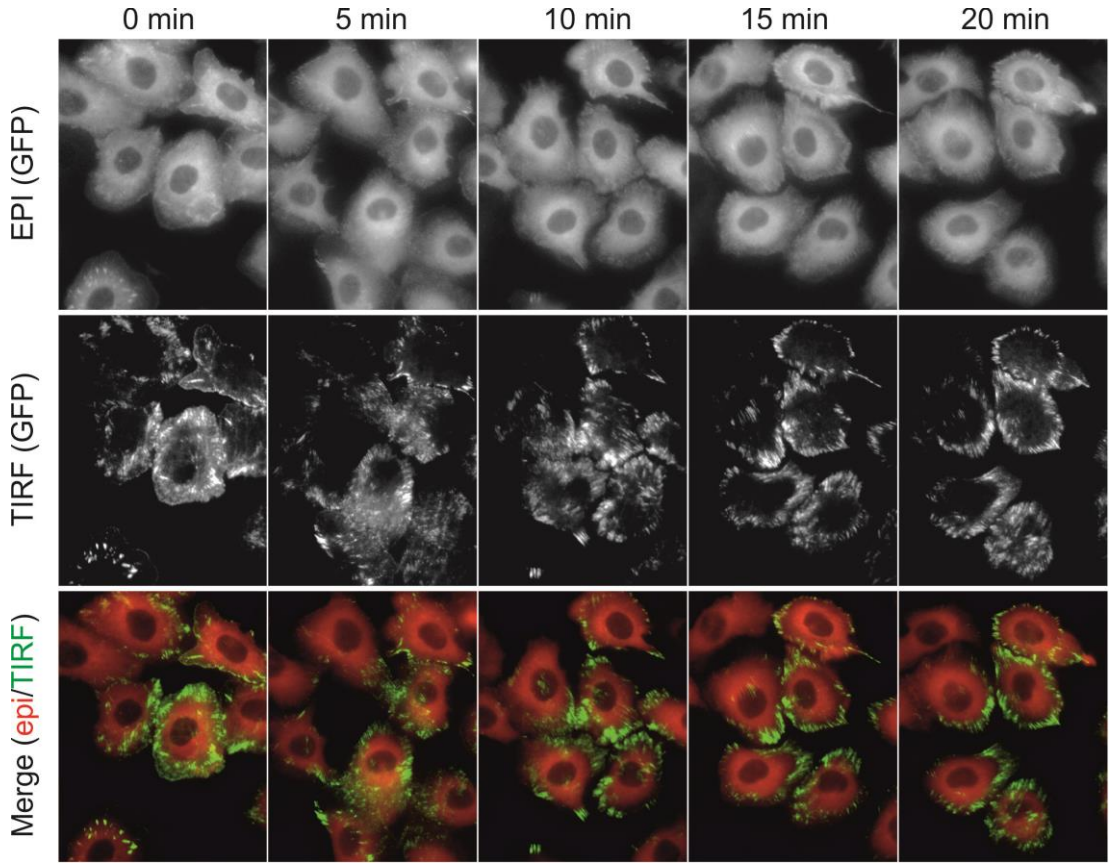


Figure 1-4 sample frames from a temporal-spatial 2D imaging setting consists of a TIRF channel and a epifluorescence channel

2. Experiment Preparation

During the experiment preparation, the researcher will conduct the experiment following the initial experiment design. Specimens, usually *in vitro* cell lines, are prepared and put into a culture plate (cf. Figure 1-5). However, the cell behavior is often subjected to environmental fluctuations such as CO_2-level, temperature [17][18], or cell-to-cell variability [19]. As a result, it is likely that an individual specimen may not homogeneously express an expected phenotypical signature. These typical wet-lab related problems are, however, beyond the scope of this thesis.

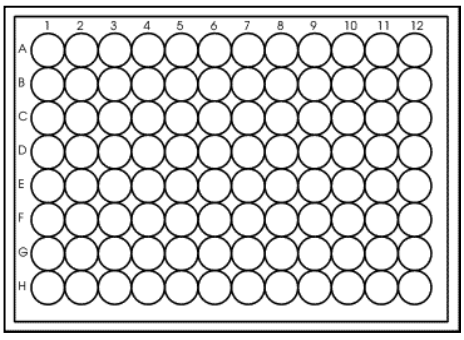


Figure 1-5 layout of a 96 well culture plate

3. Image Acquisition

In the image acquisition step, according to the plate layout, images will be produced (cf. Figure 1-6e) using microscopy imaging. Imaging techniques [10] commonly used by HT/HC screen are the following: (1) fluorescence microscopy [20][21][17][14] (cf. Figure 1-6a-c), (2) confocal laser scanning microscopy (CLSM) [22][23] (cf. Figure 1-6d), and (3) total internal reflection fluorescence (TIRF) microscopy [24][25]. When being applied for live cell imaging, the design is often referred as a time-lapse imaging (cf. Figure 1-4) for a temporal effect is studied. Similar to capturing video, time-lapse imaging captures images at a fixed sampling interval (temporal-resolution) from a predefined location in each well.

The temporal-resolution selected for time-lapse imaging, given in terms of the interval between two consecutive frames, is ranging from 30 minutes per frame for a stationary human reporting cell line [12] to 30 seconds per frame for matrix adhesion dynamics. The choice of the temporal-resolution is often considered empirical and subject-depended. The aim is to find a sample interval that captures the major changes of phenomenon. The temporal-resolution at which a phenomenon occurs usually does not require video rate. In that case, using higher temporal-resolution to capture additional intermediate stage does not necessarily provide more information. Moreover, sometimes to enforce a higher temporal-resolution may also produce undesirable quality issues such as cell death due to phototoxicity [26]. Therefore, the temporal-resolution of the imaging is tuned to take into account the preservation of cell vitality.

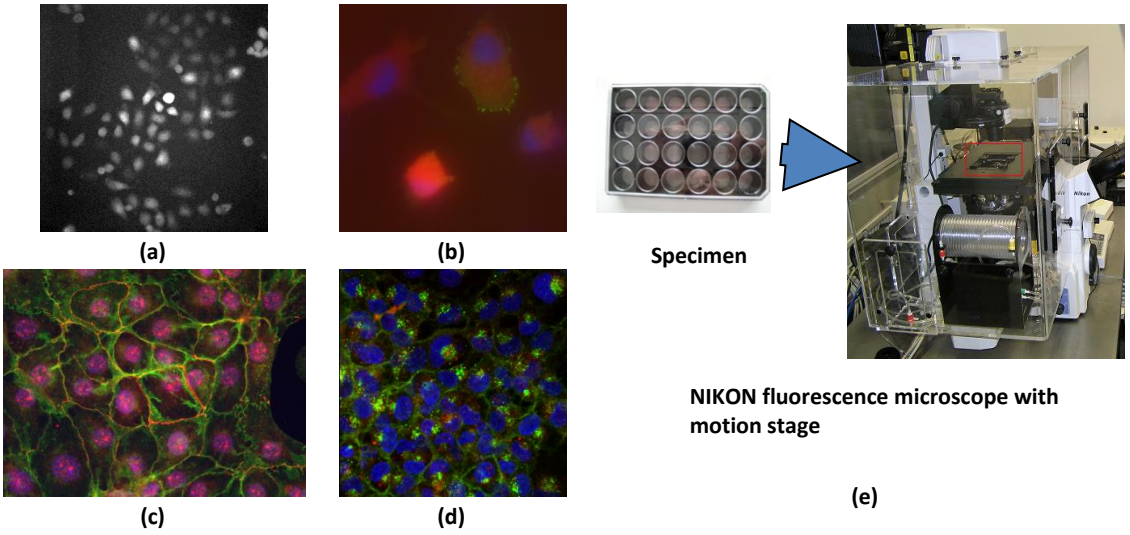


Figure 1-6 different image modalities and experiment designs in image acquisition (a) epifluorescence microscopy for cell migration analysis, (b) TIRF + epifluorescence microscopy for matrix adhesion analysis, (c) epifluorescence microscopy for protein localization during wound-&-recover, (d) confocal laser microscopy for protein translocation during cell endocytosis, (e) a simple setup of hardware for image acquisition

4. Image Analysis

The image analysis step (cf. **Chapter 2** and **Chapter 3**) is the crucial step in converting HT/HC images into numerical phenotypical descriptions. It is the bridge connecting the biological experiment and the data analysis. A robust image analysis solution is absolutely necessary to produce an objective understanding of the raw image data. In HT/HC screens, one has to deal with large quantities of images of which the quality may vary due to the fact that observations are done over a period of time as well as due to the variation in response of the cells. Consequentially, errors in image analysis will propagate into the subsequent steps. Therefore, it is important to use dedicated and problem-driven image analysis solutions. In general, our image analysis solutions for HT/HC screen studies consists of four major steps (cf. Figure 1-7): (1) **image enhancement** [27][28], (2) **image segmentation** [29][30], (3) **object tracking** [9][31], and (4) **phenotypical measurements** [10][32][33][14].

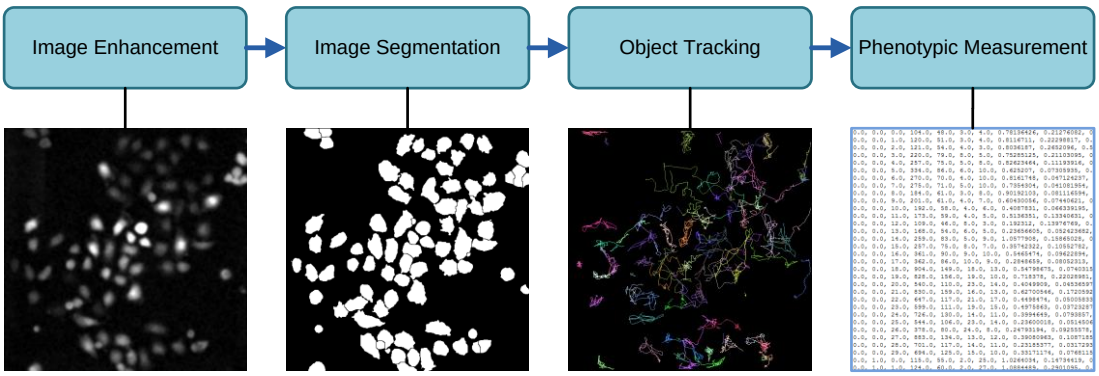


Figure 1-7 image analysis in HT/HC screen study

In the image analysis of HT/HC screen studies, the **image enhancement** serves to improve the separation between the foreground and the background. The image enhancement usually consists of two parts: (1) noise suppression and (2) signal enhancement. The noise suppression is a procedure to increase the signal homogeneity within foreground and background [34][35]. The signal enhancement is a procedure to increase the signal heterogeneity between foreground and background [34][35]. Image enhancement is not required if foreground and background can be well separated (see § 2.3). Otherwise, image enhancement is always recommended before image segmentation.

After image enhancement, the image will be segmented. **Image segmentation** is the procedure to convert the raw image into foreground and background. The foreground is defined as the pixels containing information for the analysis. In cytomics, image segmentation (cf. Ch. 2) is the essential step to convert the raw image into quantifiable elements. A good segmentation solution should correctly extract most of the foreground without introduction of background. Moreover, for HT/HC screen studies image segmentation algorithms are required to be self-adaptive and robust to image quality since it is impractical for users to tune the parameters for individual images. This is what generic segmentation algorithms often fail to comply. The image segmentation in HT/HC screen study will be further discussed in Chapter 2.

For a time-lapse imaging setting (cf. Figure 1-4), the objects from the segmented images will be tracked. The **object tracking** (cf. Ch. 3) is another crucial procedure in the image analysis to produce dynamic information of objects in a temporal study. In general, a tracking algorithm builds linkages between objects from consecutive frames using either motion models or object similarity measurements. In HT/HC screens, the efficiency of object tracking is often affected by the accuracy of image segmentation and the sampling interval of time-lapse imaging [9] [18][36]. The object tracking in HT/HC screen study will be further discussed in Chapter 3.

The phenotypic measurement is the final step of image analysis in an HT/HC screen study. It is accomplished by extracting numerical descriptors [9] using the object labeling from the image segmentation and object linkages from the object tracking. These numeric descriptors or measurements are information carried within the image data that can be used to characterize a biological phenomenon. With these measurements, data analysis can be performed to verify the initial research question.

5. Data Analysis

The data analysis step (cf. **Chapter 4** and **Chapter 5**) further augments the phenotypical measurements into comprehensive data representation and visualization (cf. Figure 1-7) by employing machine learning and statistical analysis [10][37][5][9][38][39]. The selection of data analysis solutions is depending on the experiment design. Commonly employed measures include first-order statistics, temporal low-order statistics, significance testing, data clustering and data classification. The data analysis theory is relative mature in other –omics studies, i.e. transcriptomics. In cytomics the analysis workflow is under development [29]. The decomposition and comparison of temporal data is not yet fully understood. The data analysis step is another important step, similar to image analysis, in converting image data into comprehensive conclusions.

1.3. Thesis Scope and Structure

This thesis is dedicated to the study of image analysis in HT/HC screens. A HT/HC screen produces an extensive amount of images for which manual analysis is impractical. Therefore, an automated image analysis solution should precede an objective understanding of the raw image data. The efficiency of HT/HC image analysis is often depending on image quality. Therefore, this thesis will address two major procedures in the image analysis step of HT/HC screens, namely image segmentation and object tracking. Moreover, in this thesis we will focus on extending computer science and machine learning approaches into the design of more dedicated algorithms for HT/HC image analysis. Additionally, this thesis demonstrates a practical implementation of an image and data analysis workflow case studies in which different imaging modalities and experimental designs are used. Finally, the thesis will briefly address an infrastructure for end-user interaction and data visualization.

This thesis is divided into three main parts:

Methodology

The first part includes "**Chapter 2 Robust Image Segmentation for Cytomics**" and "**Chapter 3 Robust Object Tracking for Cytomics**". It starts from a literature study of existing image segmentation and object tracking algorithms that have been frequently applied in HT/HC studies. Subsequently, it further elaborates the design and implementation of a dedicated image segmentation algorithm, namely watershed masked clustering, and two robust real-time object tracking algorithms for HT/HC studies, namely kernel density estimation with mean shift tracking and energy driven linear model tracking. In this part, the performance and efficiency of these algorithms are assessed using ground truth image sets from empirical live cell HT/HC screens. Moreover, the performance assessments of these algorithms are used to the case study part as a theoretical foundation of designing image analysis workflows for experiments.

Case Studies

The second part including "**Chapter 4 A Study to Cell Migration Analysis**" and "**Chapter 5 A Study to Dynamics of Matrix Adhesion**", demonstrates two case studies of different experiment designs and image modalities. The first case study is an analysis aiming to extract single-cell level phenotypical characterization from an aggressive cancer cell line with live cell

imaging. The second case study is matrix adhesion dynamics study. This project aims to model the correlation between the matrix adhesions, a type of subcellular macromolecular complex, and cell migration. This case study consists of a complex experiment setting and multi-modal live cell imaging.

These two case studies are demonstrating the practical implementation of image analysis and data analysis workflows following the design studies described in Chapter 2 and Chapter 3.

Supplementary System

The third part includes "**Chapter 6 HT/HC Data Management System**", illustrates a data management system to support HT/HC image analysis. Through the analysis pipeline, researchers often focus more on the image analysis and data analysis whilst an equal attention should also be given to the design of an efficient method to organize and visualize data. Moreover, HT/HC is but just one step in the cytomics pipeline. It is more important to integrate different -omics analysis to provide comprehensive understandings of biological phenomena. Thus, the design and implementation of a data management system offers a mutual platform for the data exchange between different –omics studies.

Chapter 2

Robust Image Segmentation for Cytomics

This chapter is based on the following publications

Yan, K., & Verbeek, J. F. (2012). Segmentation for High-throughput Image Analysis: Watershed Masked Clustering. *Proc. of ISoLA 2012*. LNCS 7610. Springer Berlin-Heidelberg 2012, pp. 25-41

Chapter Summary

Image segmentation is a crucial image analysis procedure in HT/HC screen study. It is often considered the first substantial step to convert image data into quantifiable elements. However, the selection of a proper segmentation algorithm is a nontrivial task. Often it requires an evaluation of both image quality and algorithm trait. To answer this question, this chapter illustrates the design of an innovative segmentation algorithm and its performance together with the assessment of several popular segmentation algorithms that have been used for HT/HC screen studies. The comparison provides an overview of the segmentation accuracy of each algorithm and a systematic assessment of the limitations of these algorithms. As a result, the comparison confirms that the watershed masked clustering (WMC), hysteresis threshold, and the level-set are good algorithms for image segmentation in HT/HC screen studies. The WMC algorithm shows the best performance with the selected HT/HC benchmark image sets.

2.1 Introduction

Image segmentation is defined as the procedure partitioning an image into multiple regions. It is often considered the most crucial step when converting image data into quantifiable elements [6][7][40][41]. Image segmentation algorithms are essentially built on two basic properties of image intensity: discontinuity and similarity. Algorithms relying on different discontinuity or similarity criteria can be further divided into the following types [42]: the edge based [43][44] and partial differential equation (PDE) [45] based algorithm are two major branches of the discontinuity-based algorithms, threshold- [46] and region-growing based algorithms are two major branches of the similarity-based algorithms.

These two branches of segmentation algorithms have proven their applicability in HT/HC screen studies. Threshold and region growing based algorithms are often employed in the image analysis of fluorescent and phase contrast microscopy due to its low computational cost and high convergence speed. However, region growing based algorithms [47] are vulnerable to local intensity variations when the size of object (foreground) is in the same range as the image noise (background).
The edge based and PDE based algorithms are used in magnetic resonance imaging (MRI) [48][49], fluorescent microscopy imaging [50][51], and phase contrast imaging. Algorithms in these two branches are considered local adaptive optimization procedures to evolve an initial curve towards the lowest potential of a cost function [52][53]. As a result, the design of a cost function often requires thorough understanding of image characteristics and cannot be easily adapted to exceptions.

Recent developments in imaging technology allow multi-channel visualization of cell structures using different microscope modalities, therefore it is important to estimate the image characteristics before selecting a proper segmentation algorithm. Our empirical study suggests that generic segmentation algorithms often proof less efficient when processing HT/HC screen images. Compared to standard imaging, quality of bio-imaging in HT/HC studies often suffers from technological complexities and experiment instabilities. In high-throughput imaging settings, these issues are the challenges to be resolved [54][55][56][57][58]. Some of the complexities may be corrected during the imaging by changing the mechanics of acquisition. Some may only be determinable after the imaging. Thus, it often relies on further digital image processing to compensate the image quality before performing segmentation. However, the processing does not necessarily guarantee an improvement in image quality.

As a result, a good selection of segmentation algorithm must be justified by heuristics; meaning to identify foreground without introducing background. Other than inventing a new principle of segmentation, here we propose a hybrid segmentation algorithm by merging existing segmentation algorithms to produce accurate masks. The **watershed masked clustering (WMC)** algorithm [9][10][29] improves segmentation efficiency by denying that all pixels may share similarity in intensity. Instead, it presumes that an image consists of a number of coarse regions in which all pixels may share the same intensity similarity. Unlike local-adaptive algorithms such as Bernsen's algorithm[59], the WMC iteratively trains only one threshold per coarse region using regional intensity values. Finally, the construction of each binary mask is tested based on morphological criteria, if necessary, a correction is imposed. By

considering both discontinuity and similarity of an image, WMC yields a significantly better performance in terms of both sensitivity and specificity. In addition, the WMC algorithm can preserve more morphological details such as cell protrusions.

In the following sections, existing HT/HC segmentation solutions together with the WMC algorithm will be introduced. Moreover, the robustness and applicability of all algorithms is estimated using ground truth image sets consisting of artificial objects and genuine HT/HC screen specimens.

2.2 Existing Segmentation Algorithms

In this section, the following algorithms are illustrated: Otsu's method [60], Bernsen's threshold [59], level-set method [45], hysteresis threshold [61], and watershed masked clustering [62]. Each algorithm is widely employed in biological image analysis. All algorithms are publicly available implementations as part of the the Fiji image analysis software [63][64]. The WMC algorithm is implemented using the ImageJ software [65].

2.2.1 Otsu's Method

In image processing, Otsu's method [60] is a threshold based segmentation algorithm that automatically performs image threshold optimization based on the histogram shape [66]. The algorithm assumes that the image contains two classes of pixels or a bi-modal histogram (e.g. foreground and background). It calculates the optimum threshold separating those two classes so that the combined spread (intra-class variance) is minimal [60]. Sometimes Otsu's method overlooks the intensity variation between individual objects (cf. Figure 2-1a). It is clear that Otsu's method can well detect objects of high intensity but cannot preserve detailed structure such as cell protrusions (cf. Figure 2-1b).

2.2.2 Bernsen's Threshold

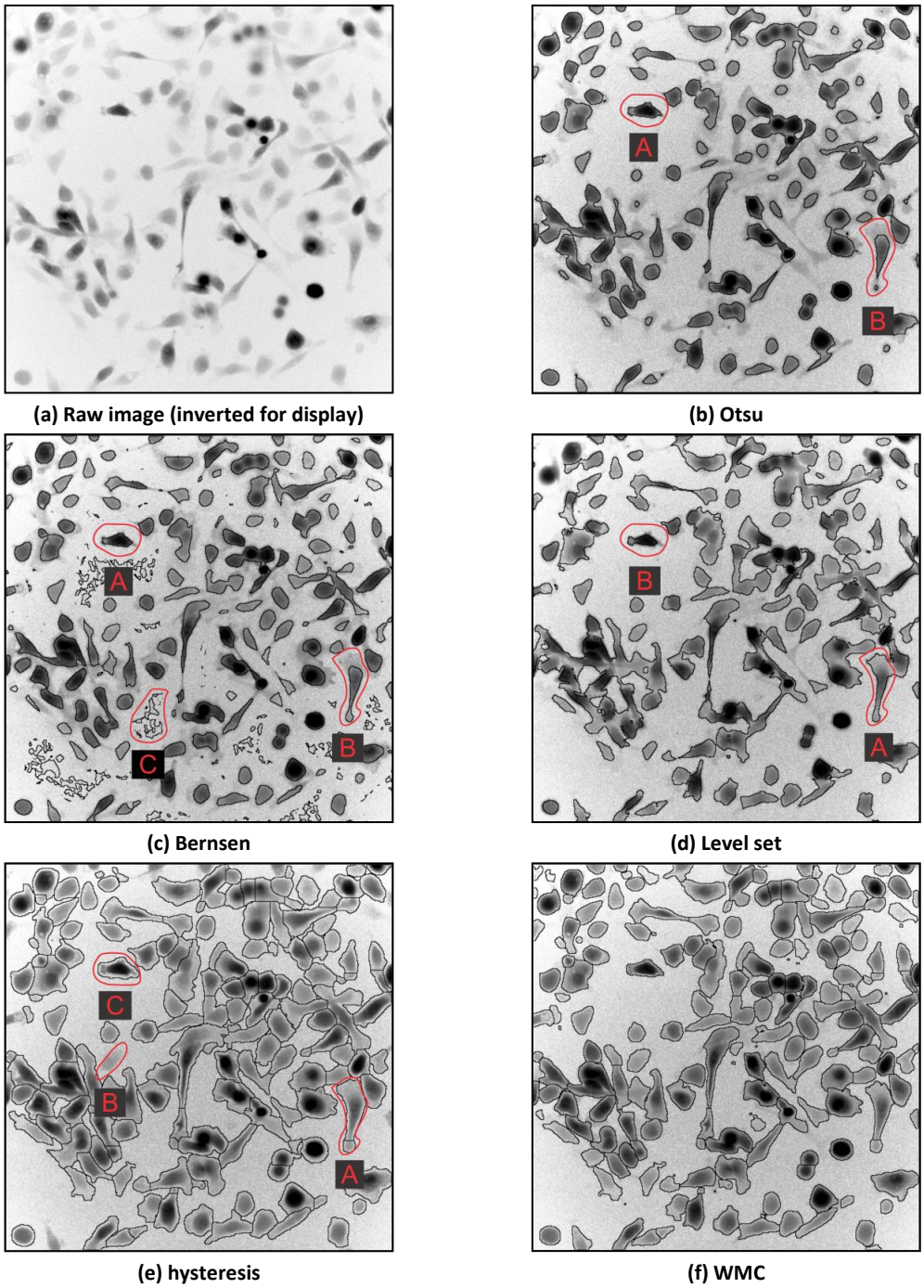
Bernsen's Threshold or Local Method of Bernsen [59] is a local adaptive threshold-based segmentation algorithm. Instead of training a threshold for the whole image, the threshold is trained for each pixel using a certain similarity of the pixel neighborhood ω [34][60]. For example, a midrange based implementation calculates the threshold using the mean of the minimum $I_{low}(i,j)$ and maximum $I_{high}(i,j)$ gray value in a local window of a predefined kernel size ω in the image *I*[60]. In each neighborhood ω, the central pixel is labeled as either foreground or background when compared to the neighborhood threshold. However, if the contrast $C(i,j) = I_{high}(i,j) - I_{low}(i,j)$ is below a certain value, then that neighborhood is assumed to consist of only one class.

Compared to global thresholding solutions, Bernsen's threshold tries to adapt the threshold for each pixel. However, The Bernsen's threshold does not consider prior probabilities of two classes in the computation of the thresholds. It also does not adapt the window size when the sizes of objects vary. Therefore, Bernsen's threshold has limitations similar to Otsu's method. (cf. Figure 2-1c).

2.2.3 Level-set Segmentation

Level-set segmentation [45][67] is a segmentation algorithm based on partial differential equation (PDE). This segmentation algorithm is derived from the level-set method, which is

originally designed for tracking objects while later adapted to the application domain of image segmentation. The central idea of such adaption is to propagate a seed or a region contour inside an image with a propagation velocity that depends on similarity measurement, such as image gradient, until a boundary (discontinuity) is reached [68].



(a) Raw image (inverted for display) (b) Otsu

(c) Bernsen (d) Level set

(e) hysteresis (f) WMC

Figure 2-1 (a) Epifluorescence microscopy images from live tumor cells. These are used for research in migratory behaviors of cancer cells. (b) – (f) Segmentation of sample images using different segmentation algorithms [A]: correctly segmented, [B]: under-segmented, [C]: over-segmented

Compared to threshold-based algorithms, the level-set considers both similarity and discontinuity of an image. As a result, it results in better recognition of object details and more robust to intensity variation (cf. Figure 2-1d). Still, the robustness of the level-set method is depending on the direction estimation and stop criterion.

2.2.4 Hysteresis Thresholding

The hysteresis thresholding [44][61] is a bi-thresholding procedure typical employed in two class segmentation problem [69]. In hysteresis thresholding, image is first segmented by an upper threshold known as “high-edge” in order to obtain only the confidence object pixels. These pixels are guaranteed to be the true foreground pixel with a higher false negative ratio. Then a second threshold is introduced as a lower threshold known as “low-edge”, which obtains probable object pixels but with a higher false positive ratio. From the “low-edge” and “high-edge” pixels, the segmentation is achieved by connecting “high-edge” pixels with “low-edge” pixels using predefined kernel; i.e. linear or quadric.

The robustness of the hysteresis method is depending on the choice of both the low-edge threshold and the high-edge threshold. Existing studies also demonstrate the possible to perform an adaptive selection of both thresholds [43]. Compared to Otsu, Bernsen, and level set, the hysteresis thresholding method yields a significantly higher masking accuracy (cf. Figure 2-1e). However, the choice of both thresholds requires constant tuning due to variation between images which is not allowed in HT/HC image study. Moreover, due to the fixed thresholds, hysteresis method tends to undertrain the masking (cf. Figure 2-1e) and object containing only weak edges is still undetectable using hysteresis thresholding (cf. Figure 2-1e).

2.2.5 Watershed Masked Clustering Algorithm

The watershed masked clustering (WMC) algorithm is designed to be a robust and dedicated solution to the application domain of HT/HC studies [10][29][9][15][70] using fluorescence microscopy. The WMC algorithm consists of three sequential steps (cf. Figure 2-2b). At each step, the segmentation result is recursively refined based on image heuristics. (1) In the first step, the WMC algorithm starts by employing a region selection mechanism to find several coarse regions considered as a rough mask that requires further optimization. (2) In the second step, a more precise masking is obtained from each coarse region using machine learning: e.g. fuzzy C-means clustering. (3) In the final step, the refined masks are reassessed based on multiple phenotypical criteria and corrected by merging, if necessary. Following this workflow (cf. Figure 2-2), the WMC algorithm converts a multimodal optimization problem into a collection of local optimization problems.

Compared to other segmentation approaches, the WMC algorithm is very robust (cf. Figure 2-1f) to regional variation of intensity in images (cf. Figure 2-1a). Moreover, unlike the hysteresis algorithm, WMC does not introduce a higher rate of false foreground when increasing its sensitivity. This trait is particular important for images with a low signal-to-background ratio. The WMC algorithm will be discussed in detail in the next section.

2.2.5.1 Region Selection

The WMC algorithm starts by defining coarse regions (cf. schema 1) using a region selection mechanism. In the current implementation, this is accomplished using the maxima-seeded watershed algorithm [30], in which the growing of the watershed region is initialized from a pixel with the highest intensity compared to its neighboring pixels: this particular pixel is referred to as the local maximum. In order to define a valid local maximum, the intensity of such a pixel must exceed the lowest pixel intensities by a threshold value *h*, where *h* is an estimated level of noise tolerance in terms of pixel intensity, *h* is commonly referred to as the

h-maximum (cf. Figure 2-3c). Higher values of *h* provide a less sensitive watershed separation and *vice versa* (cf. Figure 2-3a & b). In practice, a higher value of *h* often leads to incomplete separation of the objects in the image. Moreover, objects that occur in clusters are often not sufficiently separated (cf. Figure 2-3b). We can derive the range for the value of *h*, since the *h*-maximum is considered a reference relative to the intensity value of the pixels. Let I_M be the maximum intensity in the dynamic range of the sensor, and I_{max} the maximum intensity in the region under study, the h-maximum is typically in [*1,* (I_M-I_{max})]. In Figure 2-3, the results of the maxima-seeded watershed for different values of *h* are depicted. From empirical observations in the images typical to HT/HC experiments (I_M=*255*), a value *h*=10 provides satisfactory watershed regions.

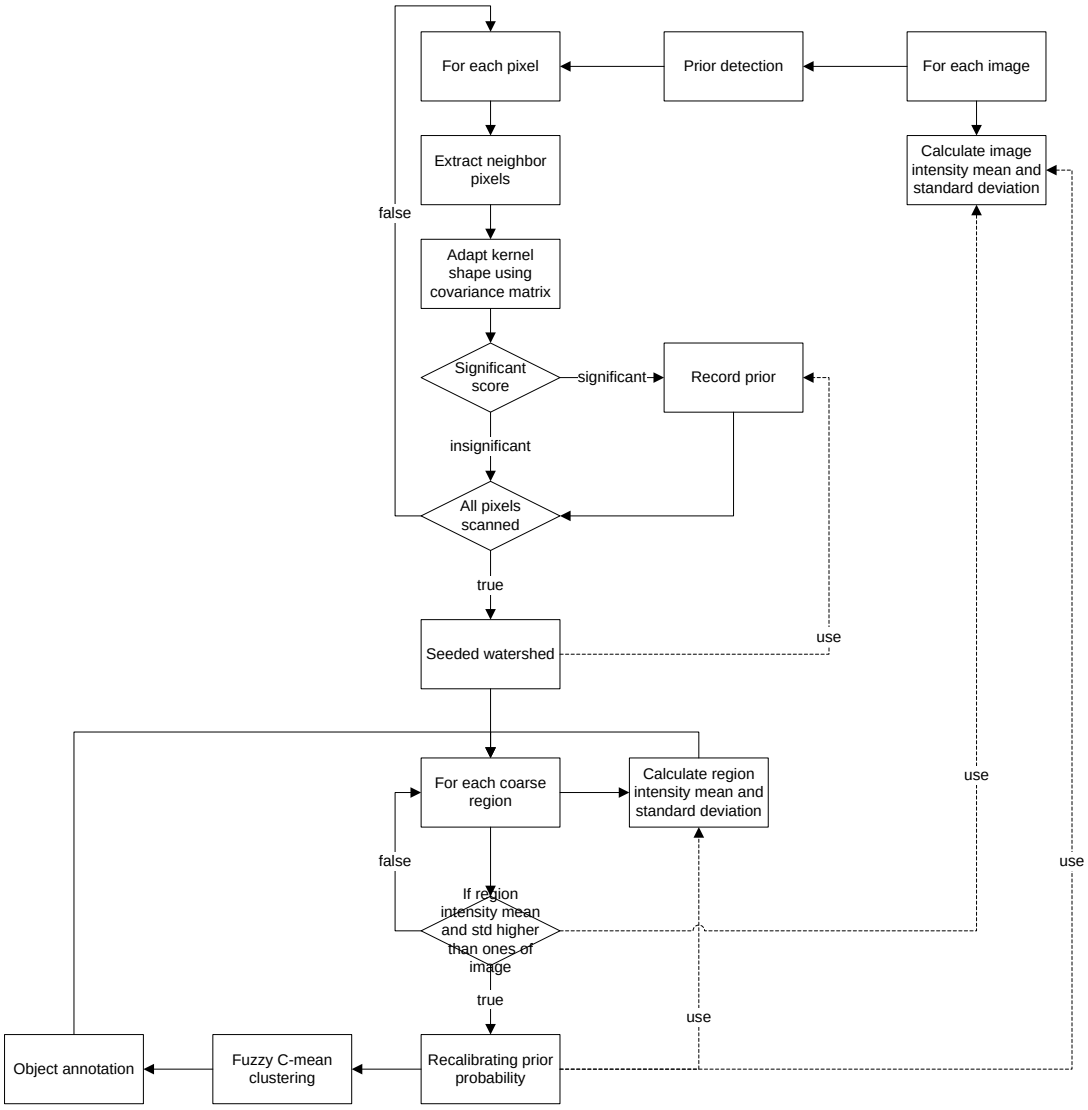


Figure 2-2 is the illustration of the three main steps of the Watershed Masked Clustering Algorithm. In the schema, the subsections are indicated in which that particular step is discussed in detail. As part of the automation process, at completion of the loop there is always a quality check Q to prevent wrongly processed images to be part of the analysis.

Given coarse regions, it is guaranteed that:

1. In each watershed region, the intensity landscape is always unimodal [30][71].
2. Seeded watershed implements a restriction on the possible starting point of path searching. An empty region usually does not contain valid seed, thus no watershed region will be formed in an empty region.

Schema 1 Watershed Masked Clustering Algorithm

```
Perform maxima-seeded watershed segmentation
Reverse watershed line into coarse region
for each coarse region r do
    False-check on coarse region
    if region is valid then
        Perform weighted fuzzy C-means clustering in intensity space I of r
        Obtain labeling
        Create regional mask
    end if
end for
Combine regional mask into final object label
```

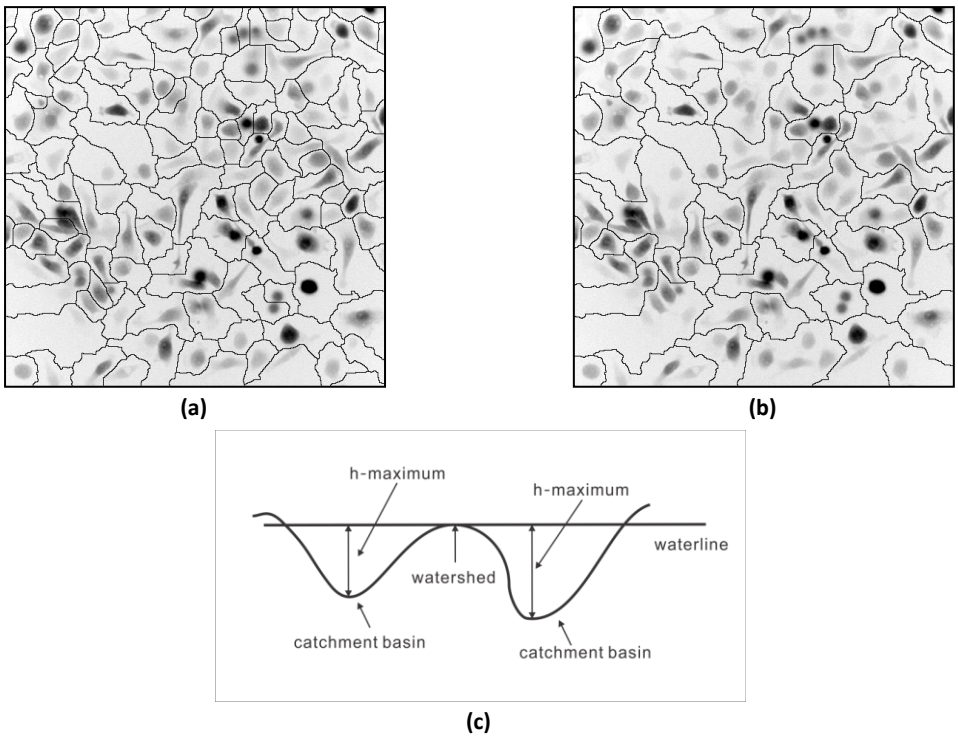


Figure 2-3: The (a) shows the definition of h-maximum. The (b) illustrates the watershed cutting lines at *h*=20. The (c) illustrates the watershed cutting lines at *h*=50.

Regarding variation in the extreme case, the result of the region selection may still contain invalid regions. Therefore, similar to Bernsen's threshold, the WMC algorithm implements a false-check mechanism for invalid regions that survive the region selection process. The false-check mechanism is constructed using intensity information including both standard deviation and mean of regional intensity values (cf. Equation 2-1). A valid candidate (coarse) region

should fulfill the first criterion, meaning the region should have large intensity diversity that suggests the presence of both sufficient foreground and background pixels. In the case that a region fails the first criterion, the second criterion will further distinguish the situation and make a final decision: i.e.

$$\begin{cases} \sigma(I_r) \geq min_{std} \\ \mu(I_r) \geq min_{mean} \end{cases} \quad \textbf{Equation 2-1}$$

, where I_r is the intensity values of one coarse region. Within these coarse regions, a threshold is trained based on local intensities. The threshold training procedure will be illustrated in the next section.

2.2.5.2 Threshold Training

With the coarse region, binary masking can be further refined from each coarse region. In order to perform such refinement, an approach is required that is capable of establishing a local adaptive threshold while being computational finite. Such can be accomplished by a weighted fuzzy C-means clustering algorithm (WFCM) [72][73]. This clustering is applied sequentially and an optimal threshold is calculated within each of the regions. Consequently, each region has its own threshold value taking into account local conditions, i.e. the local variation in image intensity.

In addition, the WFCM method has a set of weighting factors ω that allows the introduction of prior probability of the membership of the pixels in the clusters. The definition of such a weighting factor is similar to the reverse version of the prior probability in the Bayesian theorem. A smaller weighting factor is assigned to the cluster having, potentially, a larger standard deviation and vice versa. The sum of all weighting factors is always one. The weighting factor ω can be directly derived from the data [71][74], however, with a known type of image data, e.g. HT/HC images, commonly a preset value is used. The implementation of the WFCM method is described in the pseudo-code as:

Schema 2 Weighted Fuzzy C-means Clustering algorithm

```
Given weighting factor ω
Initial membership matrix u at step k=0
for each k step do
   Calculate the center vector c_j for each cluster j given Equation 2-2
   Update the membership matrix u to (k+1) given Equation 2-3
   Creating regional mask
end for
Combing regional mask into final mask
```

The WFCM method is formalized as:

$$u_{ij} = \left(\sum_{c}^{k=1} \left(\frac{\omega_j \cdot \|x_i - c_j\|}{\omega_k \cdot \|x_i - c_k\|} \right)^{\frac{2}{m-1}} \right)^{-1}, \quad \textbf{Equation 2-2}$$

where u_{ij} denotes the membership matrix, c_j is the j^{th} cluster, x_i is the data vector *i* and ω_j is the weighting factor for cluster *j*. Empirically, it was established that for cell imaging a value of $\omega = 0.2$ for the foreground and a value of $\omega = 0.8$ for the background is sufficient. This should be interpreted as: (1) there is an 80% chance a certain pixel is belongs to the foreground and (2) there is a 20% chance that a certain pixel is belongs to the background. By increasing the weighting factor for the foreground, less intense structures such as protrusions of a cell or objects with a low overall intensity will be preserved. In this manner, the weighting factor works similarly to the parameter for the degree of sensitivity in the fuzzy c-means clustering (FCM) algorithm [74]. Along with Equation 2-2, the clusters are formalized as following:

$$c_j = \frac{\sum_N^{i=1}\left(u_{ij}^m \cdot x_i\right)}{\sum_N^{i=1}\left(u_{ij}^m\right)}, \qquad \textbf{Equation 2-3}$$

where u_{ij} denotes is the membership matrix at step *k* and *m* is the, so called, fuzzy coefficient that expresses the complexity of the model, by default *m=2*. In our algorithm, we strive at a quick convergence of the WFCM and therefore the initial seeds for *c* are defined as follows:

$$c_{Foregroundseed} = \bar{I} + \left(2^{nb} - 1\right) \cdot \frac{I_{max} - \bar{I}}{\sigma(I)} \qquad \textbf{Equation 2-4}$$

$$c_{Backgroundseed} = \bar{I} - \left(2^{nb} - 1\right) \cdot \frac{\bar{I} - I_{min}}{\sigma(I)} \qquad \textbf{Equation 2-5}$$

where I_{min}, I_{max} denote the minimum/maximum intensity in the image I, $\bar{I}$denotes the mean of the intensities in image I, $\sigma(I)$ denotes the standard deviation in the intensities of the image *I* and *nb* denotes the dynamic range of the intensity expressed in number of bits. In the standard case of unsigned 8-bit images *nb=8*.

The flexibility of weighted fuzzy C-means clustering is a more robust solution when addressing the complexity in the HT/HC images. The application of this step results in a binary object in each of the regions of step 1 (cf. §2.2.5.1), if correct, shape features can be derived. However, the watershed method might have introduced some irregularities in the establishment of the coarse regions, which requires an additional evaluation; this evaluation is elaborated in the next section.

2.2.5.3 Object Optimization

At onset of our algorithm, the watershed segmentation is applied resulting in regions that are individually processed. Depending on the variation in the data, the watershed algorithm is known to result in an overcut of the segmentation; overcut is commonly referred to as the situation in which the watershed segmentation produces more regions than actually present in the image [75]. This overcut might affect the individual objects, as a result of which the objects need be split or merged (cf. Figure 2-4). Therefore, the last step in our algorithm is to compensate for the possible overcut caused by the watershed segmentation. We refer to this process as an object optimization as we evaluate the results obtained in the object segmentation. In this procedure, only the objects that share a border with a watershed line are evaluated, as these objects are the candidates for overcut. The procedure is summarized as follows:

Schema 3 Object Optimization

```
Given watershed line in step 2
for each pixel l_i in watershed line l do
  for each pixel l_n in the 4-connected neighbor of l_i do
    if pixel l_n is overlapping with the binary mask of an object then
      Object is sharing the watershed line pixel
    end if
  end for
  if pixel l_i is shared by more than two objects then
    Calculate all criteria
    if all criteria are true then
      Discard the pixel l_i
    end if
  end if
end for
Combing regional mask into final mask
```

The solution for the object optimization is a merging mechanism that uses multiple criteria; currently, two criteria are implemented but depending on the type of data; more can be added. The two criteria are:

1. Evaluation of the strength of watershed line; the objects are merged based on a local difference in maximum and average intensity in the object.
2. Evaluation of the orientation of the objects; the object are merges based on assessment of the difference in orientation of their principal axes.

For criterion 1, we implemented an intensity-based merging algorithm so as to estimate the necessity of merging the objects through the evaluation of the strength of the watershed lines. In this function all watershed lines are evaluated. This criterion can be generalized with the evaluation function *K*:

$$K\ (l_i) \rightarrow min\left(\frac{\delta_1}{\tau_1}, \frac{\delta_2}{\tau_2}\right) > T_k \qquad \textbf{Equation 2-6}$$

where the l_i denotes the ith watershed line, δ_1 denotes the difference between the average intensity under the watershed line and maximum intensity of object on one side of the watershed and similarly, δ_2 represents the object on the other side of the watershed line; where τ_1 and τ_2 denote the difference between the maximum and minimum intensity value within one object on either side of the watershed line l_i. A valid watershed line should fulfill the condition given in Equation 2-6. If $K(l_i)$ exceeds a threshold T_k then the objects on either side of the line are merged to one and the watershed is neglected. In Figure 2-4a, the intensity-based merging criterion is illustrated.

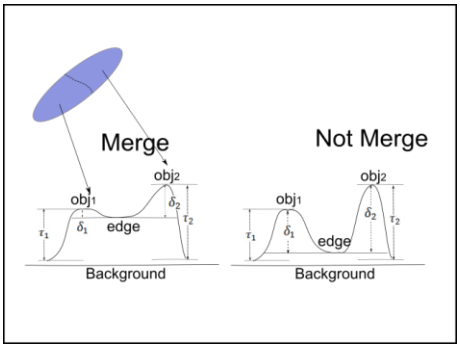
For criterion 2, we implemented an orientation-based merging algorithm [76][77][78], which provides a unique possibility to split/merge large structure complexes or elongated objects (e.g. protrusions). At watershed line l_i we consider the principal axis of the objects on either

side of the line. A two component Boolean function is designed so that when true, i.e. both components are true, the objects will be merged and the watershed line will be neglected. This function *P* is written as:

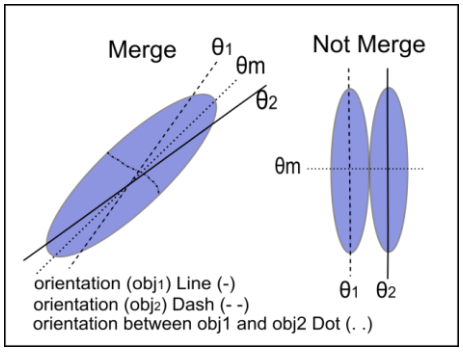
$$P(l_i) \rightarrow \begin{cases} |\theta_1 - \theta_2| < T_p \\ |\theta_1 - \theta_m| + |\theta_2 - \theta_m| < T_p \end{cases} \qquad \textbf{Equation 2-7}$$

where θ_1 denotes the angle between the horizontal image axis (x-axis) and the principle axis of object 1, similarly θ_2 is defined for the object on the other side of watershed line l_i. The θ_m is the angle between the horizontal image axis and the line crossing the centers of mass of the two objects (cf. Figure 2-4b). The components in $P(l_i)$ are separately evaluated; so, if the principle axis of each individual object spans a minimum angle T_p while the line crossing the centers of mass of the two objects lies within the angular wedge T_p of the two principle axes, only then these two objects will be merged. In Figure 2-4b, the orientation-based merging is illustrated by two cases.

In studies where the objects are cells orientation merging is used less frequent whereas in the studies on analysis of protein expression in endocytosis or cell signaling it is often applied.

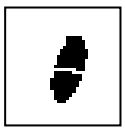


(a) Two typical cases of intensity based merging; (left) a merge is realized and (right) a merge is not realized using *K*(l_i) (cf. Equation 2-6).

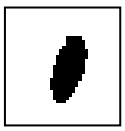


(b) Two typical examples of orientation based merging; (left) a merge is realized and (right) a merge is not realized using *P*(l_i) (cf. Equation 2-7).

(c) Sample object



(d) Overcut from step 1



(e) Merged from step 3

Figure 2-4: Illustration of the merging of objects based on a combination of criteria; in (c,d,e) a specific case for one object (cell) is illustrated.

Once the object optimization is applied, one can be certain that all objects are correctly extracted and these can be subject to a characterization of the shape. For the specific case of the time-lapse images in HT/HC, both the shape of the binary object and the intensity profile can be measured. The intensity profile of an object is derived by applying the (final) binary mask to the original image. In addition to standard features, higher order features can be used [78]. In the case of HT/HC, the features are used to discriminate between the experimental conditions that are applied [10][4]. Examples of the application of this step of the WMC

algorithm are worked out in the next section (cf. §2.3) where performance of WMC is compared to other segmentation algorithms.

2.3 Performance Evaluation

This section will address the performance assessment of the WMC segmentation algorithm together with several popular HT/HC-proven algorithms. In order to get a good impression of their robustness and accuracy, four image sets (cf. Figure 2-5) including image sets with artificially generated objects and images from genuine HT/HC experiments are produced. The ground truth image sets employed are the following:

1. Artificial image set with artificial objects (cf. Figure 2-5a)
2. HT29 phalloidin image set (cf. Figure 2-5b)
3. MTLn3 GFP image set (cf. Figure 2-5c)
4. MA image set (cf. Figure 2-5d)

The artificial image contains a number of randomly generated ellipsoid objects. The HT/HC image set is based on imaging of living migrating tumor cells or subcellular structures that ectopically express fluorescent protein [18][80]. The performance estimation for each algorithm is derived from the comparison between the binary mask obtained by the algorithm and the corresponding ground-truth binary mask for each image (cf. Figure 2-5e-h).
In the generation of the test images, the ground-truth masks for the artificial test images are explicitly constructed. The usage of such artificial image provides an image test set with an unbiased ground-truth and controllable noise, emulating a common case in fluorescence microscopy. The employment of genuine HT/HC image sets illustrates the empirical performance of the segmentation algorithm under different experiment designs or image modalities. The construction of a ground truth in these image sets are accomplished via manual delineation.

The image quality of each test set, in terms of segmentation complexity, is measured by image coefficient of variance (*image*-CV). The *image*-CV [81] is an alternative definition of the signal-to-background ratio in biomedical image processing. It is measured as the ratio between average foreground intensity and standard deviation of background intensity (cf. Equation 2-8).

$$CV = \frac{\mu_f}{\sigma_b} \qquad \textbf{Equation 2-8}$$

A higher *image-CV* suggests that the distribution of foreground intensity value is far from the distribution of background intensity value. Therefore, it measures the significance of the difference between foreground intensity and background intensity.

The image set with artificial objects (cf. Figure 2-5a) is an image set which consists of images containing artificially generated objects resembling the basic phenotype of the cells. With artificially generated objects, here we eliminate the observation bias frequently occurring in manual ground truth construction. Moreover, we use image set containing artificially generated objects to introduce hypothetical noise to test algorithm robustness. The *image*-CV (cf. Figure 2-6a) shows that the image quality of this image set is similar to HT/HC screen studies.

The HT29 phalloidin image set (cf. Figure 2-5b) [82] is a public available ground truth image sets from a HT/HC screen study. Here we choose the image channel containing phalloidin (cytoplasm staining) channel as the test images due to its complexity in term of segmentation. However, the image quality of HT29 is still higher than both artificial image set and MTLn3 GFP image set (cf. Figure 2-6a), representing the difference between foreground and background is higher in HT29 images.

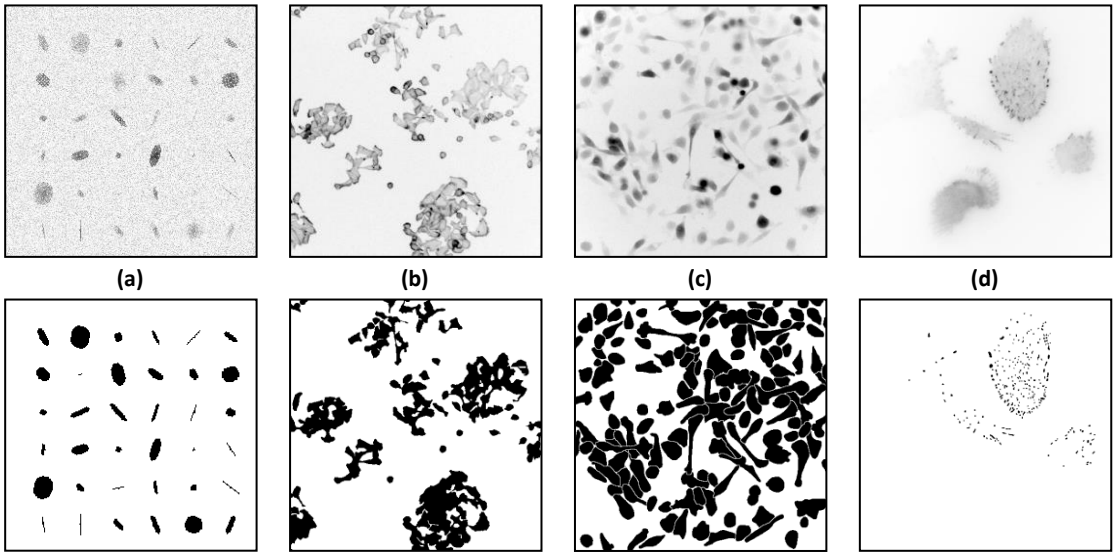


Figure 2-5 (a) artificial image set and ground truth mask (e); (b) HT29 phalloidin channel and ground truth mask (f); (c) MTLn3 GFP channel and ground truth mask (g); (d) MA image and ground truth mask (h)

The MTLn3 image set (cf. Figure 2-5c) is a ground truth image set from a HT/HC screen study using an aggressive cancer cell line. In terms of migration speed, the rat breast carcinoma MTLn3 cell-line is considered one of the most aggressive in vitro assays. The MTLn3 cells migrate as individual cells with an average velocity of 40 µm/hr [9][83] whereas the HT29 cell-line only migrates with an average velocity of 4 µm/hr. GFP signal distribution of MTLn3 cells is however more complex since the GFP protein expression is variable from cell to cell. The *image-CV* (cf. Figure 2-6a) shows that image quality of the MTLn3 GFP image set will not be easy to segment.

The MA image set (cf. Figure 2-5d) is a ground truth set consists of subcellular structure known as the matrix adhesion (MA). Due to the imaging settings, the image contains some bleed-through signal from other channels. Thus, it presents a more complex segmentation problem (cf. Figure 2-6a).

The four image sets are believed to be a reasonable representation of the quality of HT/HC images. The Figure 2-6a shows that HT29 phalloidin has the best image quality for segmentation (*image-CV*=27.85). It confirms the quality of the HT29 image set over the MTLn3 GFP image. The artificial image set has a CV similar to MTLn3, suggesting the artificial image set does emulate the real-world scenario of HT/HC screening. Among all ground truth image

set, the MTLn3 GFP image set has the worst image quality for segmentation. The MA image set shows a slightly higher (*image-CV*=8.16) than MTLn3 GFP.

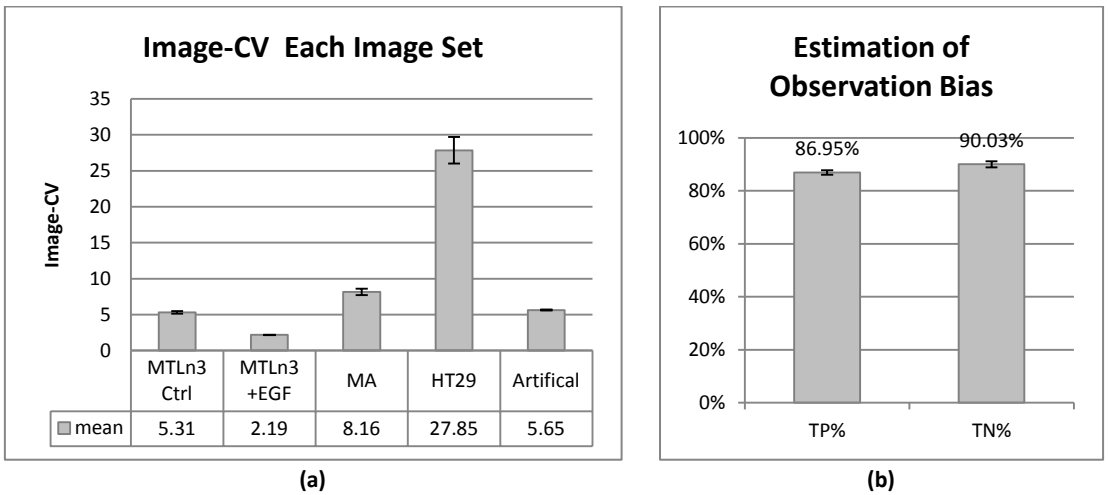


Figure 2-6 (a) *image-CV* estimation for each image set, (b) estimation of observation bias over MTLn3 GFP image set

Additionally, manual delineations are repeated with a random shuffling and rotation of the images to cover observation bias. From the estimation of the observation bias with MTLn3, it shows that on average the observers have a F1-score of 89.87% (cf. Figure 2-6b). From the observation bias, it is clear that ground truth mask is not 100% reproducible, which sustains the necessity of the artificial image set. The observation bias of HT29 is not available. With the ground truth image set, we will further introduce the assessment metrics and methodology in the next section.

2.3.1 Assessment Metrics and Methodology

In this section, the pixel-level mismatch is calculated for all segmentation algorithms. The rationale behind this test is to simulate the typical data processing workflow for HT/HC, therefore the parameters used for each of the algorithms are optimized only once and henceforth applied to the whole image set in the experiment. For none of the algorithms in the experiment an individual tuning is applied. The parameters for all algorithms (cf. § 2.2) were obtained from the high-content screening literature [14][84][17][12] and existing software [63][64][82].

Before introducing the error estimation methodology, a clarification of the assessment metrics is presented. Segmentation algorithms are often considered simplified versions of two-class classifiers that are trained in intensity space [60][59][46]. Therefore, similar to the error estimation for the classifier, the error test normally covers both type-I error (False Positive) and type-II error (False Negative). The performance of a segmentation algorithm can be assessed using the number of correct and incorrect segmented pixels [74]. This definition only covers the type-I error (FP) which may lead to an overtraining of the algorithm [85]. For a balanced conclusion, we take into account both type-I error (FP) and type-II error (FN). Furthermore, instead of just using the two errors types, we introduce the F1-score [86] which is derived from the types-I/II errors, as the major criterion of segmentation performance.

The two types of errors for different algorithms are defined in terms of the True Positive and True Negative. True positive (TP) is defined as the ratio of pixel overlap between the ground-truth mask and the segmented mask by each algorithm. This ratio is expressed as:

$$TP = \frac{M \cap M'}{M}, \qquad \textbf{Equation 2-9}$$

where *M'* is the set of pixels belonging to the foreground of binary mask provided by the algorithm and *M* is the set of pixels belonging to the foreground of the ground-truth mask. In similar fashion, the true negative (TN) is calculated as:

$$TN = \frac{\bar{M} \cap \bar{M}'}{\bar{M}}. \qquad \textbf{Equation 2-10}$$

In this way, TP represents the percentage of correctly segmented foreground pixels whereas TN represents the percentage of correctly segmented background pixels. Form the values of TP and TN, the false positives (FP) are derived, i.e. FP = 1-TP (percentage of incorrectly segmented foreground pixels), and likewise the false negatives (FN) are derived, i.e. FN = 1-TN (incorrectly segmented background pixels). From these values, the sensitivity and the specificity [87] are calculated by:

$$sensitivity = \left(\frac{TP}{TP + FN}\right) \qquad \textbf{Equation 2-11}$$

$$specificity = \left(\frac{TN}{FP + TN}\right) \qquad \textbf{Equation 2-12}$$

Given the results, the specificity and the sensitivity for all of algorithms of a particular set of test images can be computed. In addition, from the specificity and sensitivity, the F1-score is derived by:

$$F1 = 2 \cdot \frac{specificity \cdot sensitivity}{specificity + sensitivity}. \qquad \textbf{Equation 2-13}$$

A good segmentation algorithm should yield the highest F1-score but this only occurs when both specificity and sensitivity are approaching 100%. Our choice for the F1-score aims to enforce a balanced performance in terms of preventing either oversegmentation or undersegmentation.

In the next few sections, the performance assessment of the following algorithms will be exemplified. These algorithms include Bernsen local threshold algorithm, Otsu threshold algorithm, Level-set algorithm, hysteresis threshold algorithm, together with the WMC algorithm. All of the algorithms have claimed the intrinsic capacity of performing well under noisy conditions typical to HT/HC imaging [78][14][84][17][12]. For these algorithms, open-source plug-ins are available in Fiji [63][65][64] and CellProfiler [82] and these implementations were used without modifications and the tuning of the parameters for each algorithm is accomplished by cell biology specialists. These algorithms will be tested for all four ground truth image sets.

2.3.2 Artificial Objects and Test Images

In order to understand and verify the behavior and performance of segmentation algorithms without introducing observation bias, ground-truth images with objects resembling the shapes, which are normally found in high-content imaging, are constructed (cf. Figure 2-7). Each image consists of a number of ellipsoid objects and each object has a unique intensity profile. The intensity landscape is generated through an exponential decay function that is initiated at the centre of each object. The minimum and maximum value of an intensity profile of an object are generated using a uniform distributed random generator and scaled in the range of (20, 255). In this way, it resembles a random intensity variation in HT/HC screen. In addition, the orientation of each of the objects is varied by applying a rotation to each of the object. The rotation is in the range of [-30º,30º] using the center of mass as the pivot; the rotation angle is selected from a uniform random generator that is scaled to the rotation range. The random rotation of objects aims to test whether the segmentation algorithm is sensitive to direction.

The original binary image with all the objects is kept as the absolute ground-truth mask for the segmentation so that error estimation can be applied over a range of test images. In this test, a total amount of 30 images is generated. To simulate image noise typical to HT/HC and fluorescence microscopy, Poisson noise is generated by giving each pixel a random intensity (Poisson) [35][88] oscillation. In the noise-added images, the intensity oscillation is ranged from 0 to 255 while the size is approximately 3 to 5 pixels. In the original image, the object intensity will be above zero and therefore a true global threshold can always be found.

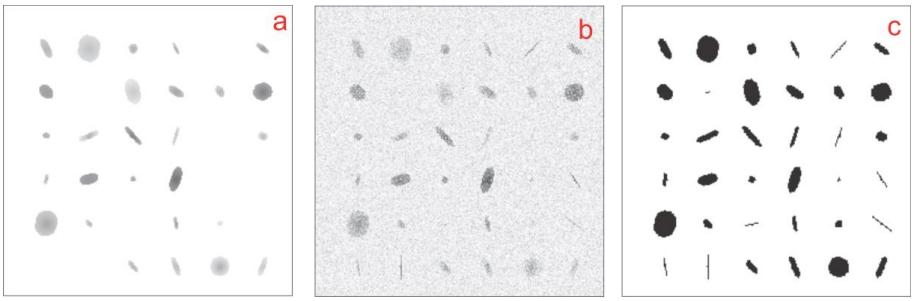
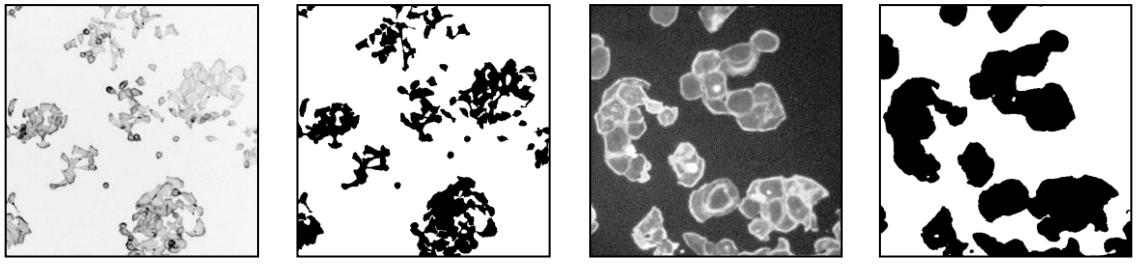


Figure 2-7: Artificial image set being used for assessing the efficiency of the segmentation. (a) Original Image in 256x256 (b) Noise-added Test Image (c) Ground-truth Mask

2.3.3 HT29 Phalloidin Images

The HT29 phalloidin image set is known as the “Human HT29 Colon Cancer” dataset [84] (cf. Figure 2-8); a publically available image dataset including ground-truth image set provided by the Broad Institute of MIT. This set contains 12 images of human HT29 colon cancer cells (cf. Figure 2-9a & c). Each image consists of three channels including phalloidin (channel 1), pH3 (channel 2), and nucleus (channel 3 Hoechst). Hoechst labels DNA present in the nucleus. Phalloidin labels actin present in the cytoplasm. The pH3 stain indicates cells that are in division and will not be used in this test. For the benchmark study, we focus on image segmentation performance with the phalloidin channel (cf. Figure 2-8c) because it contains several complexities including intensity variation, uneven illumination and phenotypical variation in object size (cf. Figure 2-8a & c).



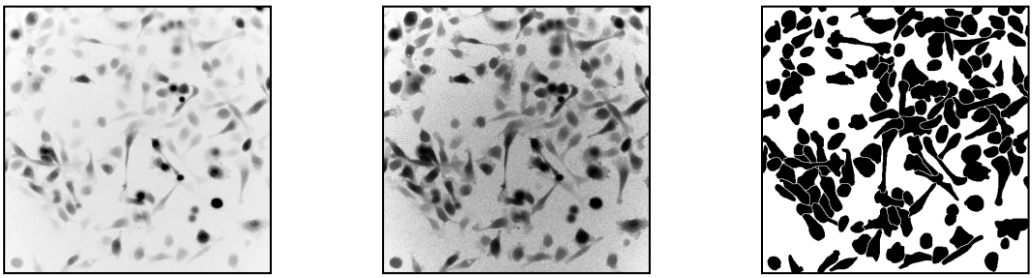
(a) Original for (b) (b) Ground-truth mask (c) Original for (d) (d) Ground-truth mask

Figure 2-8: images (512x512) from the HT29 phalloidin channel, which (a) represents phenotype of control cells, which (c) represents phenotype of treated cells. There is an observable size variation between (a) and (c).

2.3.4 MTLn3 GFP Images

The MTLn3 GFP image set is a time-lapse image sequence, i.e. a dynamic process, of the breast carcinoma MLTn3 line used to understand tumor cell migration in the context of breast cancer metastasis (cf. Figure 2-9). The set is provided by the Leiden Academic Center for Drug Research (LACDR). It consists of 96 time-lapse image sequences, each of 75 frames in 5 minute sampling intervals. Each sequence portrays an *in vitro* cell migration pattern typical in HT/HC experiments. The GFP (Green Fluorescent Protein), was ectopically expressed to label the cell body (cytoplasm + nucleus), enabling fast fluorescent imaging [80]. For the performance tests, we will only use the first 14 images of the sequence to reduce the size of the image set to reasonable for proportions for this test. In addition, for this image set also a ground-truth image is required. The MTLn3 ground-truth images were obtained by manual segmentation performed by biologists through tracing on a digitizer tablet (WACOM, Cintiq LCD-tablet) (cf. Figure 2-9c). In contrast to the artificial image set, manual segmentation may contain bias between and within observers. To that end, the manual segmentation is replicated a few times. Moreover, to improve visibility of objects in each image, the observer will draw the cells using images to which an intensity equalization is applied [89] (cf. Figure 2-9b).

The MTLn3 set is a good representation of a high throughput screen with *in vitro* live cell migration (cf. Figure 2-14) for it contains intensity variation, uneven illumination, phenotypical variation and object overlapping.



(a) MTLn3 gray image (b) Image (a) with equalization (c) MTLn3 ground-truth mask

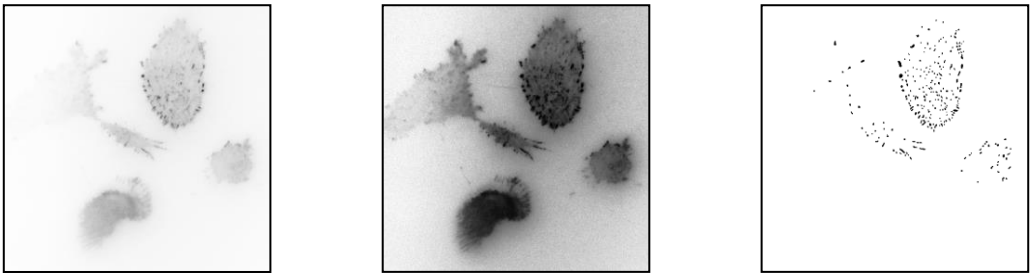
Figure 2-9: Typical image (512x512) from the MTLn3 GFP set

2.3.5 MA Images

The MA image set is a time-lapse image sequence of matrix adhesion (MA) dynamics captured using total internal reflection fluorescence (TIRF) microscopy [25]. This set is provided by the Leiden Academic Center of Drug Research as a study of correlation between matrix adhesion regulation and cell migration. The MA image set consists of subcellular macromolecular complexes visualized using TIRF microscopy (cf. Figure 2-10). The magnification of MA TIRF is

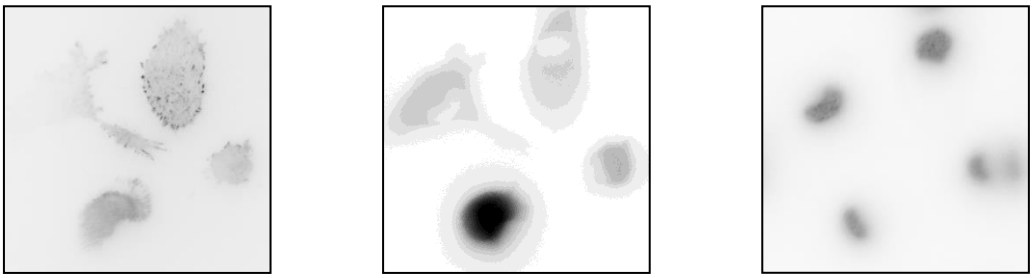
higher than other sets but the object size is relatively small (approx. 5~65 pixels). Due to the smaller size, each object contains less intensity information for the segmentation algorithm. In terms of ground truth construction, observers reported difficulty in distinguishing between background and foreground. The procedure for the ground truth set is the same as the MTLn3 image set.

The MA image set is included as a case study for algorithms' extensibility to both relatively small objects (MA) and relatively large objects (cell body) (cf. Figure 2-11 a & b). Here we define relatively small objects as objects occupying less than 0.01% of whole image and the relatively large objects as objects occupying more than 1% of image. For a 512x512 image, it means a small object will be approximately 10~50 pixels while a large object will be approximately 5000~10000 pixels.



(a) MA gray image **(b) Image (a) with equalization** **(c) MA ground-truth mask**

Figure 2-10: Typical image (512x512) from the MA TIRF set



(a) MA channel (GFP + TIRF) **(b) cell body channel (GFP + EPI)** **(c) nucleus channel (Hoechst + EPI)**

Figure 2-11 complete design of the imaging acquisition for MA image set

2.3.6 Result of Benchmark Study

Artificial Objects and Test Images

All algorithms are applied over the same 30 test images (cf. Figure 2-12). In the Table 2-1, the results for the true-positives are listed. In the Table 2-2, the results for the true-negatives are listed. The F1-scores are listed in the Table 2-3. The object merging accuracy in WMC is also tested using the same image set. An overcut object is defined as a group of objects obtained by segmentation algorithm share the same object in ground truth mask. A total amount of 238 overcut objects are detected in this image set. Using object optimization, the WMC recovers 202 out of 238 overcut objects, i.e. approximately 85%.

From Table 2-3 we can conclude that the WMC yields the highest performance while Otsu, Bernsen and Hysteresis produce acceptable results. The level-set shows the lowest performance. From the image (cf. Figure 2-12), it is noticeable that none of the segmentation algorithms has successfully captured objects smaller than noise. The WMC can well preserve shape of object regardless of noise and intensity variation, but it has a tendency of

undertraining the threshold in some cases. Both Bernsen and Otsu can also preserve objects with higher intensity homogeneity but cannot adapt to intensity variation. The Hysteresis algorithm can well preserve objects with higher intensity homogeneity but cannot preserve objects with an extreme morphology, i.e. an elongated object or a small spherical object. The level-set algorithm fails to adapt its propagation due to the intensity variation.

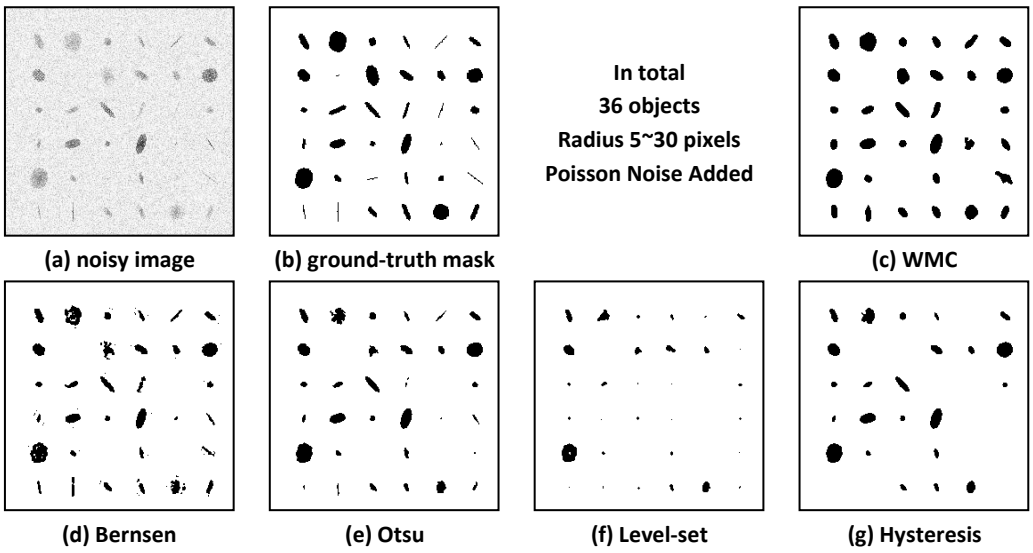


Figure 2-12: (a) noise-added test image, (b) ground-truth masks for the object, (c) to (g) are binary images obtained by corresponding segmentation algorithms.

Table 2-1 True positive rate

True Positive	WMC	Bernsen	Otsu	Level-set	Hysteresis
Avg	96.67%	91.83%	98.48%	60.57%	86.39%
Std	3.81%	6.23%	1.68%	28.14%	2.74%

Table 2-2: True negative rate

True Negative	WMC	Bernsen	Otsu	Level-set	Hysteresis
Avg	95.34%	74.73%	84.00%	99.01%	78.10%
Std	5.45%	28.36%	18.46%	1.36%	12.64%

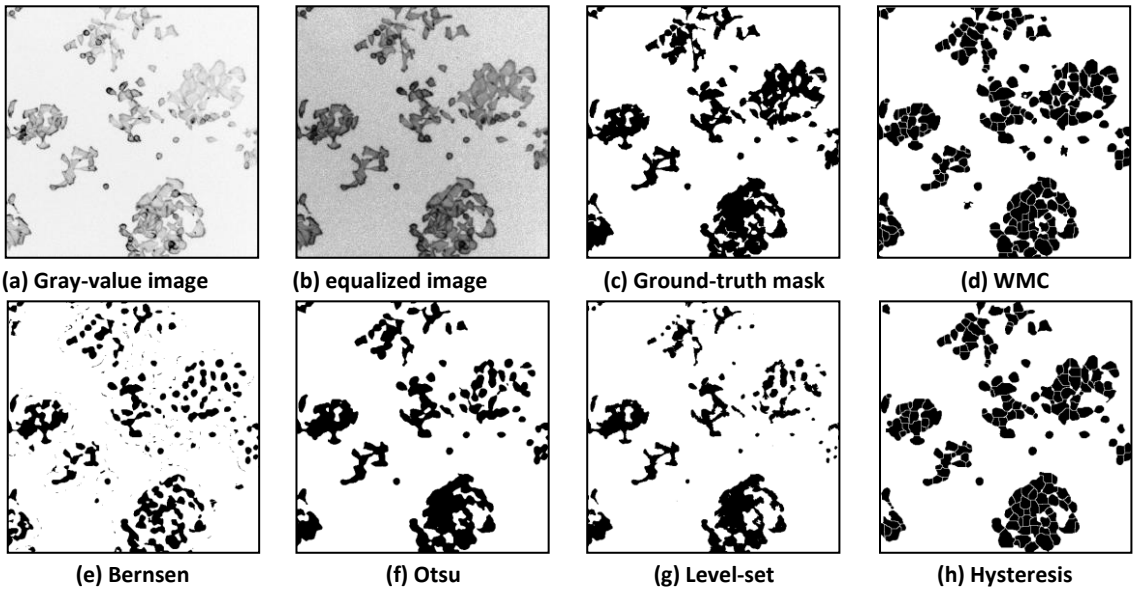
Table 2-3: Specificity and sensitivity of segmentation efficiency using artificial images

Performance	WMC	Bernsen	Otsu	Level-set	Hysteresis
Sensitivity	96.62%	90.14%	98.22%	71.52%	85.16%
Specificity	95.40%	78.42%	86.02%	98.38%	79.78%
F1 Score	95.99%	84.60%	91.83%	74.98%	82.95%

HT29 Phalloidin Images

From Table 2-4, we can conclude that nearly all algorithms can perform well with HT29 image set (cf. Figure 2-13d). It is unclear which algorithm is better since WMC, Otsu, and Hysteresis are all producing good results. In general, WMC (cf. Figure 2-13d), Hysteresis (cf. Figure 2-13h) and level-set method (cf. Figure 2-13g) are all performing well with the raw images. The Otsu segmentation result (cf. Figure 2-13f) shows that a global threshold algorithm is less robust to intensity variation but its performance can be improved by intensity equalization. The halo

structures in Bernsen (cf. Figure 2-13e) is believed to be associated with a localization of staining at the cell border region, which is smaller than the kernel size chosen in Bernsen.



(a) Gray-value image (b) equalized image (c) Ground-truth mask (d) WMC

(e) Bernsen (f) Otsu (g) Level-set (h) Hysteresis

Figure 2-13: (a) Original HT29 image acquired with a 10x lens; image size is 512x512 pixels, 8 bit. b) the image (a) with intensity equalization c) Ground-truth masks and (d-h) masks obtained by the segmentation algorithms.

Table 2-4: Specificity and sensitivity of the segmentation algorithms in the HT29 set

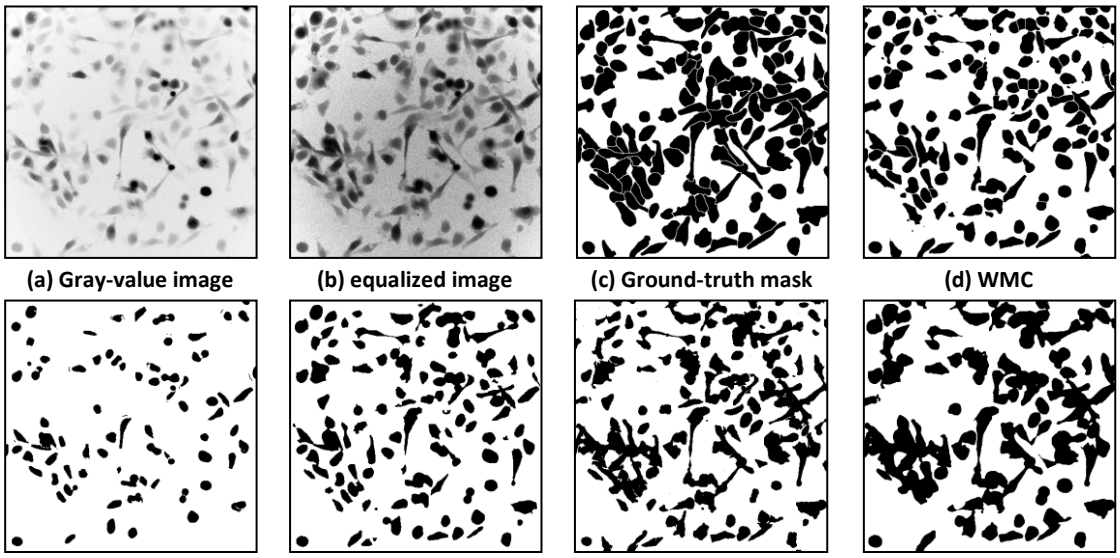
No Intensity Equalization					
Performance	WMC	Bernsen	Otsu	Level-set	Hysteresis
Sensitivity	92.49%	95.49%	99.13%	98.24%	97.51%
Specificity	91.71%	63.71%	79.48%	84.57%	83.29%
F1 Score	92.10%	76.43%	88.23%	90.89%	89.84%
Intensity Equalization					
Performance	WMC	Bernsen	Otsu	Level-set	Hysteresis
Sensitivity	85.63%	86.16%	97.06%	80.27%	84.70%
Specificity	88.21%	66.43%	93.81%	79.04%	82.59%
F1 Score	86.90%	75.02%	95.41%	79.65%	83.63%

MTLn3 GFP Images

From Table 2-5, we conclude that MTLn3 is a difficult image set for image segmentation. It is immediately clear that the overall performance is much lower compared to the experiment with the HT29 set (cf. Table 2-4). Most methods are able to extract the brighter region around the nucleus (cf. Figure 2-9) while the fuzzier cytoplasm region is not detected. With MTLn3 set, the WMC algorithm still shows the highest performance. The Hysteresis thresholding algorithm is able to portray a good and stable performance. Compared to the previous experiment (cf. Table 2-4), the performance of each algorithm is decreased. The WMC algorithm, however, performs quite stable under these different circumstances.

Although WMC and Hysteresis are able to preserve the fine details with comparable F1-score, the Hysteresis demonstrated an unbalanced performance by yielding a low sensitivity but high specificity, meaning it misses many foreground pixels. The WMC, on the other hand, shows a

more balanced performance in both sensitivity and specificity. All algorithms show improvements after intensity equalization.



(a) Gray-value image (b) equalized image (c) Ground-truth mask (d) WMC

(e) Bernsen (f) Otsu (g) Level-set (h) Hysteresis

Figure 2-14: (a) Original MTLn3 image acquired with a 20x lens (NA 1.4), image size 512x512 pixels, 8-bit. (b) the image (a) with intensity equalization. (c) ground-truth masks and (d-h) masks obtained by the segmentation algorithms.

Table 2-5: Specificity and sensitivity of segmentation algorithms in MTLn3 image set

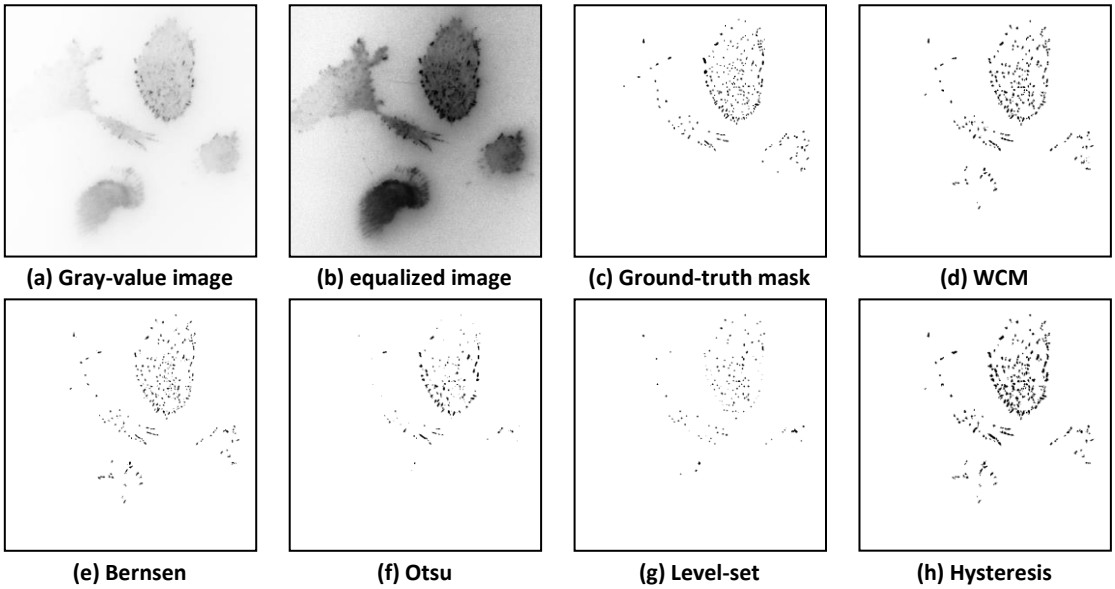
No Intensity Equalization					
Performance	WMC	Bernsen	Otsu	Level-set	Hysteresis
Sensitivity	73.02%	21.50%	17.25%	27.86%	66.86%
Specificity	90.75%	99.62%	99.67%	99.13%	94.21%
F1 Score	80.92%	35.36%	29.42%	43.50%	78.21%
Intensity Equalization					
Performance	WMC	Bernsen	Otsu	Level-set	Hysteresis
Sensitivity	85.70%	36.03%	39.17%	30.37%	79.94%
Specificity	82.54%	98.35%	98.80%	99.20%	87.25%
F1 Score	84.09%	52.75%	56.09%	46.50%	83.43%

MA Images

The performance assessment (cf. Table 2-6) shows that WMC demonstrates a stable and robust performance similar to the previous image sets (cf. Table 2-4 and Table 2-5). This holds also for the hysteresis thresholding and level-set algorithm. Both Bernsen and Otsu algorithm show a lower performance. The smaller object size poses difficulty for all algorithms since there is less intensity information available for the threshold training or contour propagation. This has been foreseen in the test with artificial objects (cf. Table 2-3).

Table 2-6 Specificity and sensitivity of segmentation algorithms in MA TIRF image set

No Intensity Equalization					
Performance	WMC	Bernsen	Otsu	Level-set	Hysteresis
Sensitivity	84.00%	67.16%	62.20%	74.85%	80.70%
Specificity	97.21%	99.09%	99.74%	99.84%	98.91%
F1 Score	90.12%	80.06%	76.62%	85.56%	88.88%



(a) Gray-value image (b) equalized image (c) Ground-truth mask (d) WCM

(e) Bernsen (f) Otsu (g) Level-set (h) Hysteresis

Figure 2-15 (a) Original MTLn3 image acquired with a 20x lens (NA 1.4), image size 512x512 pixels, 8-bit. b) the image (a) with intensity equalization. (c) ground-truth masks and (d-h) masks obtained by the segmentation algorithms.

2.4 Computation Complexity

To further verify the high-throughput nature of the WMC algorithm, its computational complexity is studied and compared to the same selection of segmentation algorithms in previous sections (§2.3). The computational complexity of the WMC algorithm is derived from its two major components, namely watershed segmentation and fuzzy C-means clustering. Both algorithms are known to be NP-complete problems. In an NP-complete problem, the computational cost grows in polynomial order, but the total amount of computational cost is nondeterministic. We will therefore focus on the computational complexity in fashion of a single-iteration. In the watershed segmentation (flooding based), worst case, the searching of all descending paths is in the order of $O(n^2)$ [71].

In iteration of fuzzy C-means (FCM) clustering, the computations are divided into three steps and the total complexity is a joint sum of all individual steps.

Step 1: update membership matrix

$$u_{ij} = \left(\sum_{c}^{k=1} \left(\frac{\omega_j \cdot \|x_i - c_j\|}{\omega_k \cdot \|x_i - c_k\|} \right)^{\frac{2}{m-1}} \right)^{-1} \in O\left(n^{\frac{4}{m-1}}\right) \qquad \textbf{Equation 2-14}$$

Step 2: calculate new seeds

$$c_j = \frac{\sum_N^{i=1}\left(u_{ij}^m \cdot x_i\right)}{\sum_N^{i=1}\left(u_{ij}^m\right)} \in O(n) \qquad \textbf{Equation 2-15}$$

Step 3: validation

$$\left\|u^{k+1} - u^k\right\| \in O(n^2) \qquad \textbf{Equation 2-16}$$

The total complexity of FCM clustering in a single iteration is $O\left(n + n^{\frac{4}{m-1}} + n^2\right)$. Together with the computational complexity of watershed segmentation, the total complexity of WMC can be written into $O\left(n + n^{\frac{4}{m-1}} + n^2\right)$. With fuzzy factor *m=2*, the computational complexity of the WMC is in the order of n^4. The practical computational performance of each algorithm is given in the Table 2-7. The performance is based on a test with 14 images of the MTLn3 set.

Table 2-7: Computational performance in seconds; executed on a 2.4 GHz P4 (single-thread) with 4GB RAM. All implementations of the segmentation algorithms are in Java.

average time (sec)	WMC	Bernsen	Otsu	Level-set	Hysteresis
mean	4.46	754.3	1.14	2.34	0.06
std	0.22	7.58	0.02	0.17	0

From this analysis, we can deduce that the computational load of the algorithm will be in reasonable bounds and suitable for the type of application it was initially designed. Further implementation of concurrent computation in WMC has significantly reduced the computation time (cf. Table 2-8) by a linear factor equals to the number of threads.

Table 2-8: Computational cost in seconds; executed on a 2.7 GHz i7 (8-thread) with 8GB RAM. All implementations of the segmentation algorithms are in Java.

average time (sec)	WMC (multi-thread)	WMC (single-thread)
mean	0.42	3.37
std	0.08	0.12

2.5 Conclusion and Discussion

This chapter illustrated a number of image segmentation algorithms, which are applied in the image analysis procedures for high-content screens. In conclusion, from Table 2-3, Table 2-4, Table 2-5, and Table 2-6, it can be established that the WMC algorithm outperforms the other algorithms regardless image modality and image quality issues. Moreover, WMC has demonstrated a stable and robust performance for each image sets compare to other algorithms. This is in particular important for HT/HC screen studies since it is impossible to predict the output of treatments. Therefore, WMC algorithm is considered a good solution for HT/HC screen studies.

The major advantage of the WMC algorithm is that it can deal with variations in staining intensity typical for bio-imaging and specific to high-throughput in vitro experiments (cf. Figure 2-1). The local intensity variations in the image limit application of Otsu segmentation; it requires a global optimum for the threshold, which may not be possible. Along the same line, the level-set method is not suitable as it presumes a consistent intensity for the objects in the image. The regional approach in WMC followed by a local clustering transforms the segmentation to a local problem so that threshold levels can be found efficiently. For segmentation in cytomics edge based methods are noise susceptible, therefore intensity variations necessitate region based approaches. This is confirmed from our findings comparing Hysteresis segmentation to WMC, especially with more artificial noise or staining variations in the image (cf. Figure 2-1).

The WMC consists of three independent steps and if we consider these individually further improvements can be formulated. In step 1, the watershed algorithm, the initialization of the watershed algorithm is currently based on local maxima; other schemas must be investigated to render a better initialization. Now, a priori knowledge is not used whereas this might facilitate a better estimate for the initialization. In step 2, fuzzy weighted C-means clustering is used, however, other clustering approaches can be probed; similarly to step 1, a priori knowledge on the intensity distribution might be supportive in finding a better clustering approach. Regarding step 3, we implemented only a few of the situations of oversegmentation. This particular step of the algorithm can be adapted to experimental conditions, i.e. a priori knowledge can be tuned with respect to the experiment so as to overcome certain imperfections of earlier steps. In future research this will be elaborated, however, the global idea of the WMC algorithm will stand its case.

The WMC has been successfully applied to other experiments in the domain of bio-imaging, e.g. detection of small vessels [18], cell membrane [70] and cytoskeleton formation [90]. With further generalization, the algorithm can be engaged in a broader scale of imagery. The future research on the tuning of the subsequent steps of the WMC algorithm will contribute to this generalization.

Chapter 3

Robust Object Tracking for Cytomics

This chapter is based on the following publications

Yan, K., LeDévédec, S., van de Water, B., Verbeek, F.J. (2009) "Cell Tracking and Data Analysis of in vitro Tumour Cells from Time-Lapse Image Sequences". *In Proc. of International Conference on Computer Vision Theory and Application (VISAPP2009)*.

Yan, K., Le Dévédec, S., Van de Water, B., & Verbeek, F. J., "Automated Analysis of Matrix Adhesion Dynamics in Migrating Tumor Cells", (in preparation)

Chapter Summary

Object tracking or video tracking is the procedure of following moving objects over consecutive frames in a video. The main principle of a tracking algorithm is to associate target objects from consecutive frames based on given linkage criteria such as minimum shape change or motion model. When performing dynamic analysis with a live cell HT/HC screen, object tracking is essential to provide phenotypical quantifications on migration behavior. Unlike standard video recording, HT/HC live imaging is based on time-lapse microscopy with a lower temporal-resolution of only one frame for every couple of minutes. In HT/HC live imaging, the quantity, in terms of number of time points, must often be scarified to guarantee the image quality. As a result, the selection of a robust object tracking algorithm is important in the dynamic analysis HT/HC live imaging.

This chapter introduces four tracking algorithms generally applied in HT/HC live cell screens using time-lapse imaging. These algorithms are divided into **confidence measurement based** and **motion-model based**. The confidence measurement based tracking algorithms include the **blob tracking** algorithm and the **kernel density estimation (KDE) with mean shift** tracking algorithm. These assume, regardless of migration, a minimum shape change only occurs between consecutive objects. Thus, often a shape model (KDE) or an overlap measurement (blob) are employed as linkage criteria. The **motion model** tracking algorithm assumes that the object is moving in a quantifiable probability model and by knowing the previous location of the object it is possible to predict the next position of the object. Typical motion model tracking algorithms are the **particle filter** tracking algorithm employing Brownian motion model and the **energy driven linear (EDL) model** tracking algorithm employing linear motion model. The confidence measurement tracking algorithms should not be mixed with the motion model tracking algorithms since the former is an analysis based on shape similarity while the latter is an analysis relying on motion patterns.

3.1. Introduction

Object tracking or video tracking is the image analysis procedure concerned with the linkage of objects. Here we define a video is a collection of frame captured at a temporal-resolution higher than 25 frame/second while a time-lapse image sequence is a type of video captured at a temporal-resolution lower than 25 frame/second. A HT/HC live cell screens in Cytomics often produces time-lapse image sequence instead of video.

Unlike object tracking with standard video, object tracking with time-lapse image sequence faces several complexities; amongst which the velocity-to-temporal-resolution ratio (VTR) is the most essential one. We define the VTR as the ratio between object's real velocity and temporal resolution of the video. A high VTR suggests there are more variations between objects from consecutive frames. In a high VTR image sequence, the between-frame association of the recognized object can be difficult. Moreover, motion of objects in live cell HT/HC screen are often the result of nonlinear deformation [91][92] that cannot be described by a rigid model [93][94][95]. The deformations introduce more morphological variation between objects. To guarantee the reliability of the tracking, the VTR of a time-lapse image sequence must exceed a minimum threshold that captures major morphological changes in objects.

There are generally two sides to object tracking with time-lapse image sequences: (1) the recognition of the relevant objects and (2) the between-frame association of the recognized object. Recognition of the relevant objects is accomplished using image segmentation (cf. Ch. 2). It is a laborious process due to the large volume of image data and its complexity is often determined by the robustness of image segmentation techniques. The tracking algorithms are based on, but not limited to, the following two properties:

1. confidence measurement [9][10]
2. motion models [31][17]

The property (1) includes shape similarity measurements such as kernel density estimation [9][96] while the property (2) includes probability functions such as Brownian motion model [97][98]. In general, confidence measurement based tracking algorithm such as blob and mean shift tracking algorithms assume that a minimum shape change will only occur between consecutive objects while the motion model based tracking algorithms, such as particular filter tracking and energy driven tracking, assume that the next location can be predicted from all previous locations of the object. Recent studies show that further optimization based on tree diagrams [99] can improve the tracking decision.

Tracking algorithms frequently used in HT/HC live cell screen are blob tracking, mean shift tracking, active contour tracking, and particle filter tracking:

1. The blob tracking is a feature point based tracking algorithm that constructs object linkage using confidence estimation. It is a computationally cheap tracking algorithm that works well real-time tracking problems.
2. The mean shift tracking is a model based tracking algorithm based on confidence estimation. Unlike blob tracking, the mean shift tracking algorithm recursively shifts the initial model to the most likely region in the consecutive frames.

3. Active contour tracking is a model based tracking algorithm that propagates the initial contour to each consecutive frame. It is a computationally intensive algorithm that requires human interference.
4. The particle filter tracking is a feature point based tracking algorithm similar to blob tracking.

In this chapter, all tracking algorithms i.e. **blob tracking**, **KDE mean shift tracking**, **particle filter tracking**, **energy driven linear model tracking**, as well as **active contour tracking** will be explained first. Subsequently, we perform a systematic estimation of tracking efficiency within the application domain of HT/HC studies using manually produced ground truth data sets. Finally, we will demonstrate the motivation behind selection of tracking algorithms in different combinations of imaging and research in cell biology.

Blob Tracking Algorithm

Blob tracking [100][101][102][97][33] is a straightforward tracking algorithm that links objects by confidence measurement such as measuring distance [103][104] or size changes between the target object and candidate object from consecutive frames (cf. Figure 3-1). However, as it is based on confidence measurement concepts, blob tracking can only track object with a mostly rigid body transformation[103].

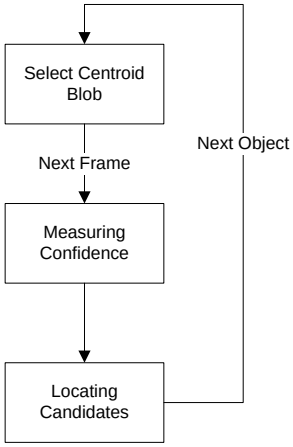


Figure 3-1 simple workflow of the blob tracking algorithm

Mean Shift Based Tracking Algorithm

The mean shift algorithm [105] is considered a real world application of an model based localization approach. It is a robust tracking solution [99][106][107][9][108][15] to associate object linkage by localizing an initial model in consecutive frames. For one n-frame video, the mean shift algorithm starts by converting the initial object into multiparametric density models and recursively update the mean shift factor based on a local density in the consecutive frames until a stationary location is reached. Finally, it associates the initial object with the candidates closest to the station location (cf. Figure 3-2). A kernel based mean shift tracking consists of two steps[105]:

1. Non-parametric density estimation from an initial model.
2. Steepest descent to locate the local maximum in a gradient space of density estimations given an initial model.

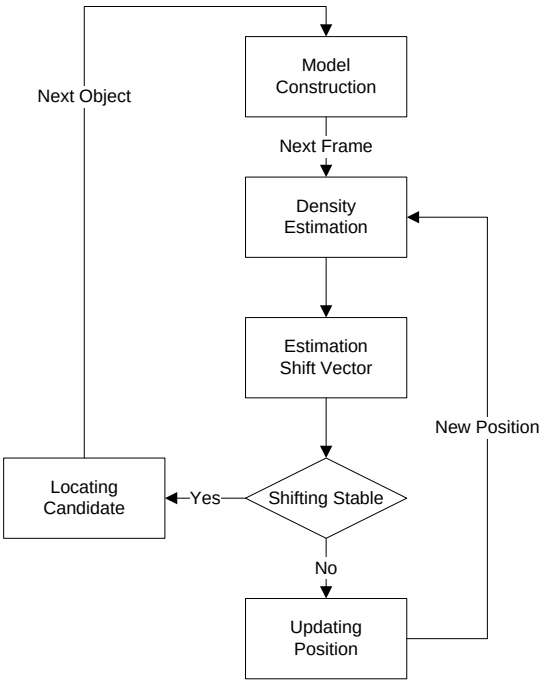


Figure 3-2 simple workflow of the mean shift based tracking algorithm

Each trajectory begins with objects in one frame. These objects are converted into initial model defined in a multidimensional feature space. These dimensions include (1) the *x*-coordinate of a binary mask of an object, (2) the *y*-coordinate of a binary mask of an object, (3) the intensity value at each pixel (*x*, *y*). Given *n* data points x_n in the d-dimensional space R^d, the kernel density estimator with kernel function *K(x)* (cf. Equation 3-2) and window bandwidth *h*, can be expressed as:

$$\hat{f}_{h,k}(x) = \frac{c_{k,d}}{nh^d}\sum_{i=1}^{n} K\left(\left\|\frac{x - x_i}{h}\right\|^2\right) \qquad \textbf{Equation 3-1}$$

A Gaussian kernel is used as a radial symmetric kernel, expressed as:

$$K_N(x) = (2\pi)^{-\frac{d}{2}}|\sum|^{-\frac{1}{2}}e^{-\frac{1}{2}\Sigma^{-1}\|x\|^2} \qquad \textbf{Equation 3-2}$$

Subsequently, the mean shift estimation is completed by steepest descent through iterative computation of:

- the mean shift vector $m\left(k^{th}\right)$
- the shifted model by $x^{k+1} = x^k + m\left(k^{th}\right)$

The steepest descent requires estimation of the gradient space $g(x) = -k'(x)$, where the mean shift vector $m\left(k^{th}\right)$ is calculated by $\frac{\sum_{i=1}^{n} x_i g\left(\left\|\frac{x-x_i}{h}\right\|^2\right)}{\sum_{i=1}^{n} g\left(\left\|\frac{x-x_i}{h}\right\|^2\right)} - x$. Because of the shape change (deformation) of objects, the steepest descent does not necessarily converge at the centre of mass of the true candidate. We chose the object closest to the stationary point, at which the magnitude of $\left(k^{th}\right)$ is closest to zero, as the probable candidate.

KDE mean shift tracking cannot provide a precise localization of the candidate position. However, it is more robust and flexible since the multiparametric density model can be derived

from any available information. As a result, the mean shift based algorithm does not necessarily rely on the relative position of candidates. Therefore, it is a good solution for tracking motile objects in time-lapse image sequences of low temporal-resolution.

Particle Filter Based Tracking Algorithm

Particle filter based tracking [36][65][109] is a lineage linkage based tracking method frequently employed in the study of random particle activity in physics [110]. This is a category of tracking algorithms that relies on a series of observations containing stochastic variation over the temporal dimension so as to produce a recursive statistical approximation; i.e., the spreading function of the underlying system states (cf. Figure 3-3). To further improve tracking accuracy, particle filter based tracking often includes linking strategies such as graph-based optimization [98] or Bayesian recursive estimation [110]. Particle filter based algorithms have been used in cell biology tracking problems like tracking random cell migration at low magnification [103].

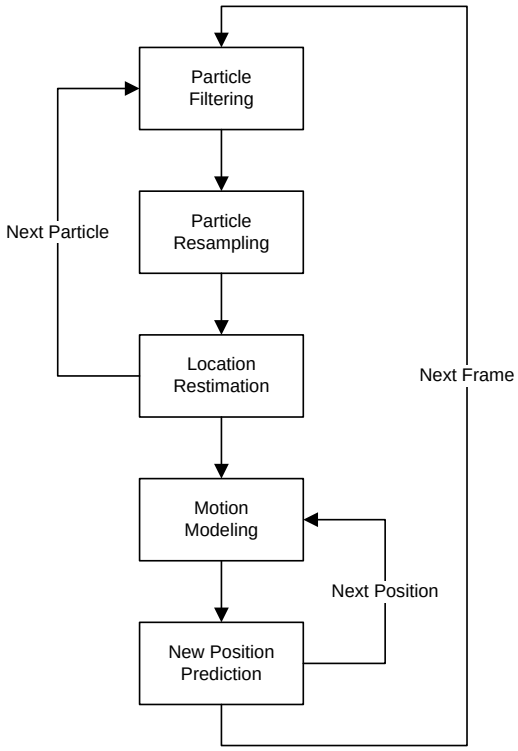


Figure 3-3 Simple workflow of particle filter based tracking

Particle filter tracking method using Bayesian recursive estimation assumes that input x_k and the observations y_k can be modeled in this form:

- $x_0, x_1, \ldots x_m$ is a first order Markov process such that $x_k|x_{k-1} \rightarrow p_{x_k|x_{k-1}}(x_k|x_{k-1})$ at state *k* and with an initial distribution $p(x_0)$.
- The observation $y = \{y_0, y_1, \ldots y_n\}$ are conditionally independent provided that x_0, $x_1, \ldots x_m$ are known. So each y_k only depends on x_k, $y_k|x_k \rightarrow p_{y_k|x_k}(y_k|x_k)$.

The linkage is constructed based on recursively updating of the posterior distribution over the current state x_t given all observations $Y = \{Y_0, Y_1, \dots Y_p\}$ up to a discrete frame number t as follows.

$$p(x_t|Y_t) = kp(x_t|y_t) \int_{x_{t-1}} p(x_t|x_{t-1}) p(x_{t-1}|Y_{t-1}) \qquad \textbf{Equation 3-3}$$

where the likelihood $p(x_t|y_t)$ at time t expresses the measurement model and $p(x_t|x_{t-1})$ is the motion model. The posterior probability $p(x_{t-1}|Y_{t-1})$ is approximated recursively as a set of weights of N samples $\left\{x_{t-1}^{(r)}, \pi_{t-1}^{(r)}\right\}_{r=1}^{N}$, where $\pi_{t-1}^{(r)}$ is the weight for the particle $x_{t-1}^{(r)}$. Using Monte Carlo approximation, Equation 3-3 can be calculated using Equation 3-4, whereas each particle is scored according to the approximation (cf. Equation 3-4).

$$p(x_t|Y_t) \approx kp(x_t|y_t) \sum_r \pi_{t-1}^{(r)} p\left(x_t \left| x_{t-1}^{(r)} \right.\right) \qquad \textbf{Equation 3-4}$$

Energy Driven Linear Tracking Algorithm

The energy driven linear tracking algorithm (EDL) [16] is a particle filter tracking algorithm that we designed for our tracking problems in subcellular structures known as the matrix adhesion (MA). The EDL is based on empirical observation of MA dynamics [36][111][112] from which MAs are believed to move in a linear fashion along stress fibers. The EDL tracking algorithm consists of the following steps:

1. Multivariate Gaussian model construction
2. Density estimation
3. Pairwise linkage
4. Trajectory construction

The complete workflow is illustrated in the Figure 3-4. We will first construct a probability model that describes the pseudo-motion behavior. Next, we denote X_n as the pixels set of the binary mask of one object at time point *n*.

For one X_n, the pixel coordinates of its binary mask are stored as a $(N \times 2)$ matrix. Then the center of mass set $\mu_n = E(X_n)$ is first calculated and next from X_n and μ_n, the covariance matrix S_n is calculated as:

$$S_n = \frac{(X_n - \mu_n)(X_n - \mu_n)^T}{N - 1} \qquad \textbf{Equation 3-5}$$

Given object μ_n and S_n, the algorithm further calculates the new center of nass $\mu_n{}'$ by $\mu_n{}' = \mu_n + V_s$, where the shifting vector V_s is calculated as $V_s = \mu_n - \mu_{n-1}$, where the μ_{n-1} is the mass center of X_{n-1}. For each X_n, an updated multivariate Gaussian model $f(X, \mu_n{}', S_n)$ is constructed based on S_n and $\mu_n{}'$ as:

$$f(X, \mu_n{}', S_n) = \frac{1}{(2\pi)^{\frac{k}{2}} det(S_n)^{\frac{1}{2}}} exp\left(\frac{-\frac{1}{2}(X - \mu_n{}')^T (X - \mu_n{}')}{S_n}\right) \qquad \textbf{Equation 3-6}$$

, where the k is the column rank of X_n; which in this case $k = 2$ since there are two columns.

Next for all X_{n+1} within the maximum searching radius r_{max}, the average per-pixel probability of X_{n+1} is calculated as:

$$p(X_{n+1}) = \frac{1}{N_{n+1}} \sum f(X_{n+1}, \mu_n', S_n) \qquad \textbf{Equation 3-7}$$

The average per-pixel probability *p* of X_{n+1} is employed as the score of the linkage. Figure 3-5 illustrates a sample result of the score table. Instead of looking at a local optimum of linkage, the pairwise linkage intends to construct a linkage of global optima between time point *n* and *n+1*. If X_n can be linked to two successors, the algorithm always selects the pair with the larger per-pixel probability given $f(X, \mu_n', S_n)$.

Finally, the algorithm searches through all pairs and constructs a trajectory. If X_n has no valid successor, the trajectory will be terminated. For this we use the following termination function *F*:

$$F(X_{n+1}) = \begin{cases} p(X_{n+1}) & \geq \ p_{min} \\ \|\mu_n - \mu_{n+1}\| & \leq \ r_{max} \end{cases} \qquad \textbf{Equation 3-8}$$

, where the termination criterion r_{max} is an user-defined maximum search radius and the termination criterion p_{min} is an user-defined minimum per-pixel probability.

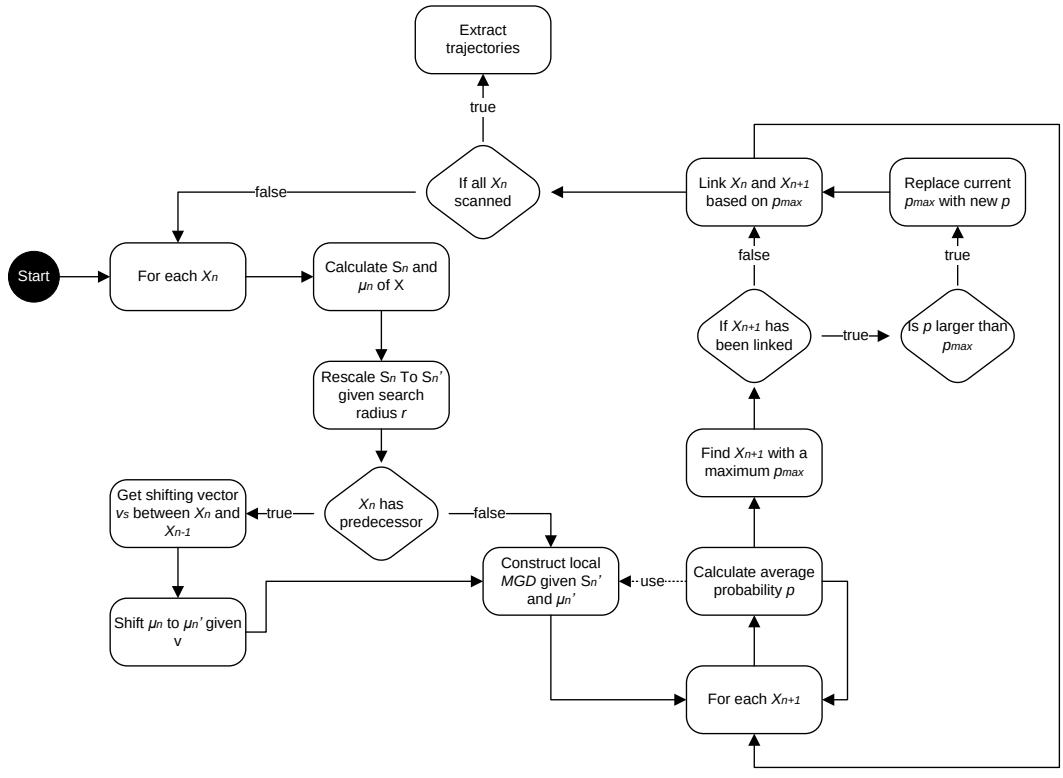


Figure 3-4 workflow of energy driven linear tracking

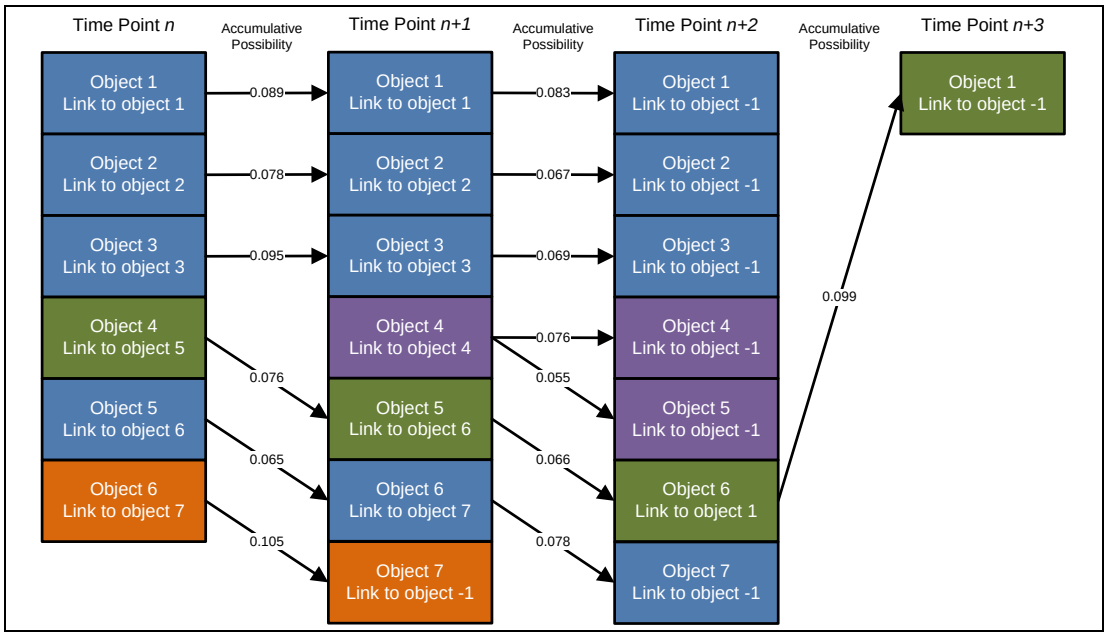


Figure 3-5 score table and pairwise linkage schema

Active Contour Tracking Algorithm

The active contour algorithm [101][103][45] is a top-down tracking algorithm combining both segmentation and tracking. From an initial model, active contour based tracking attempts to minimize an objective function associated with the current model in a recursive fashion. For an n-frame video, the active contour algorithm first adapts the initial contour for the object in frame *i* and then uses the adapted contour as the initial contour for frame *i+1* until it reaches frame *n*. Such recursive propagation of initial model not only provides a self-adaptive segmentation procedure, but also an association between objects in consecutive frames. The accuracy and robustness of active contour based tracking is subject to the choice of energy definition and local search mechanism; these must be changed for different sets of image sequences. Moreover, active contour tracking always requires a manual initialization. Therefore, the active contour based tracking is less applicable in high-throughput analysis. We included a demo tracking using active contour tracking, but did not extend this to a large scale test.

We selected an open-source implementation for each tracking algorithm category since these open-source implementations have already been used and optimized in the application domain of HT/HC studies. Using open-source implementations provides an independent starting point for the assessment of the performance of each algorithm.

- Blob tracking → MTrack2 plug-in in Fiji [64]
- Mean shift tracking → KDE mean shift plug-in in ImageJ [9]
- Particle filter tracking → particle detector and tracking plug-in in Fiji [109]
- Active contour tracking algorithm → AB snake in ImageJ [65]

These algorithms are frequently employed in HT/HC studies. The efficiency of these tracking algorithms has not yet been assessed. In the next section, we perform a systematic quantification and characterization of the tracking performance of each algorithm in the application domain of HT/HC live cell screens.

3.2. Performance Study

The tracking efficiency of each algorithm is measured from a ground truth time-lapse image sequence containing manually segmented and tracked objects. Using manually segmented objects (cf. Figure 3-6c), we can assess the tracking accuracy of each algorithm without considering possible segmentation errors. Moreover, since each algorithm uses different information for linkage calculation, the manually segmented objects provide a mutual starting point by restrict tracking to be performed only with the manually segmented objects.

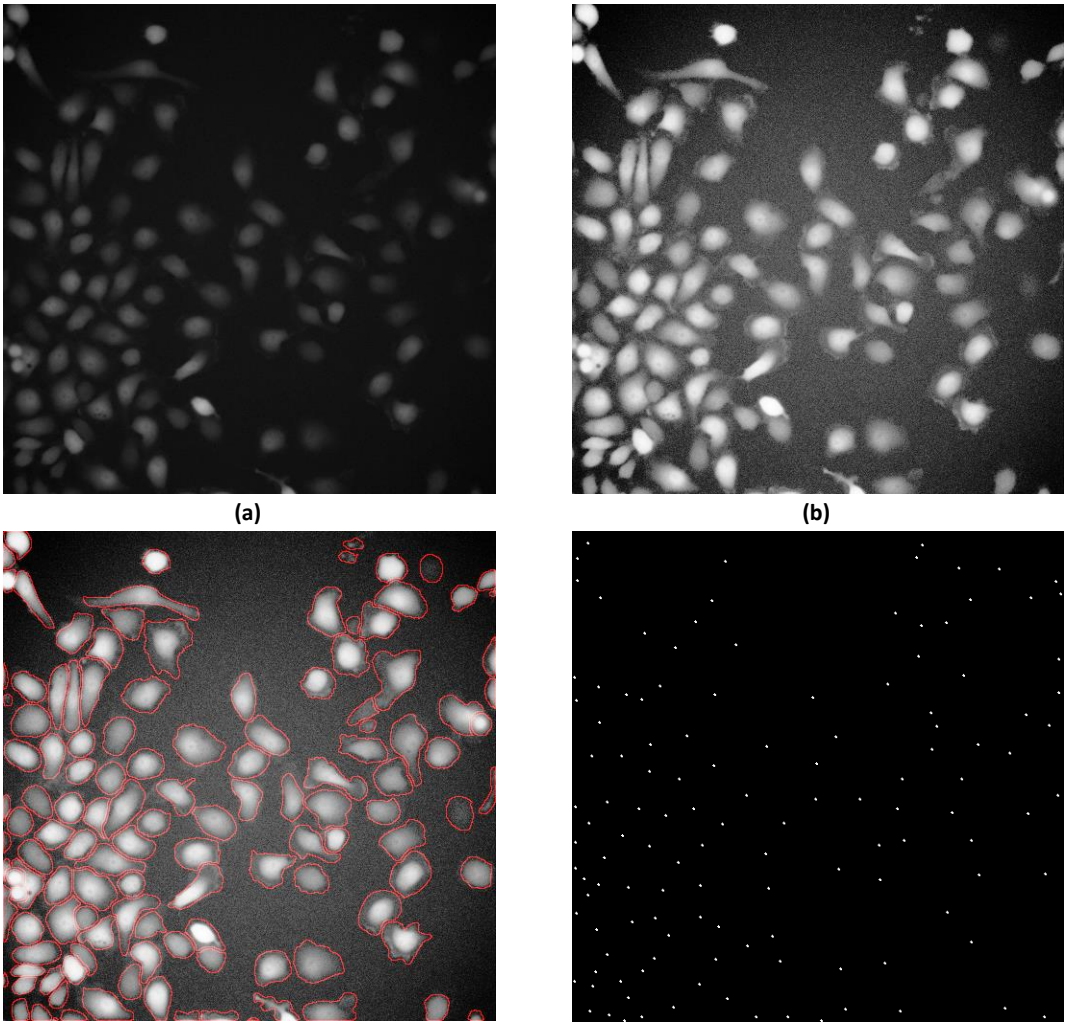


Figure 3-6 (a) raw image from the HT/HC screen of random cell migration, (b) intensity equalized version of raw image (a), (c) manually segmented and tracking cells, (d) mass center of each object

To assess tracking efficiency, here we introduce an assessment metrics consisting of five elements [113]: (1) true positive tracking, (2) false positive tracking, (3) global detection rate, (4) global fragmentation rate and (5) final score. The procedure is depicted in Figure 3-7 and described as following:

Step 1 Ground Truth Production

Observer is required to produce binary masks and trajectories of dynamic objects based on given time-lapse image sequence.

Step 2 Tracking with Prior Object Annotation

The manually segmented binary mask (cf. Figure 3-6c) will be used by the blob and mean shift tracking algorithms. The manually segmented binary mask is converted into centers of mass for particle detection and tracking (cf. Figure 3-6d). In theory, such design should guarantee that all tracking algorithms are tracking the same set of objects.

Step 3 Converting Tracking Output to Mutual Ground

Each trajectory is converted into a multi-dimensional array (cf. Table 3-1) in which the *path#* is the unique index of each trajectory, the *frame#* is the frame index, the (*MC_X*, *MC_Y*) is the (*x*, *y*) coordinate of the center of mass the object. Given one trajectory, the extraction of corresponding object masks is illustrated in Figure 3-9. Since the different algorithm implementations produce various formats of output to describe trajectory information, the only possible way to compare the tracked trajectories from different implementation is to map trajectories back to the objects.

Step 4 True Positive Tracking (TP) and False Positive Tracking (FP)

Using state definition in Figure 3-8, the true positive tracking and false positive tracking are calculated for each tracking algorithm.

Step 5 Global Detection Rate (GDR)

Using the result from true positive tracking and false positive tracking, each path tracked is assigned to one ground truth path. From the global coverage, the global detection rate is calculated.

Step 6 Global Fragmentation Rate (GFR)

Using fragmentation definition in Figure 3-8, the global fragmentation rate is calculated for each tracking algorithm.

Step 7 Final Tracking Score

The final tracking score is derived from TP, FP, GDR and GFR (cf. Equation 3-16).

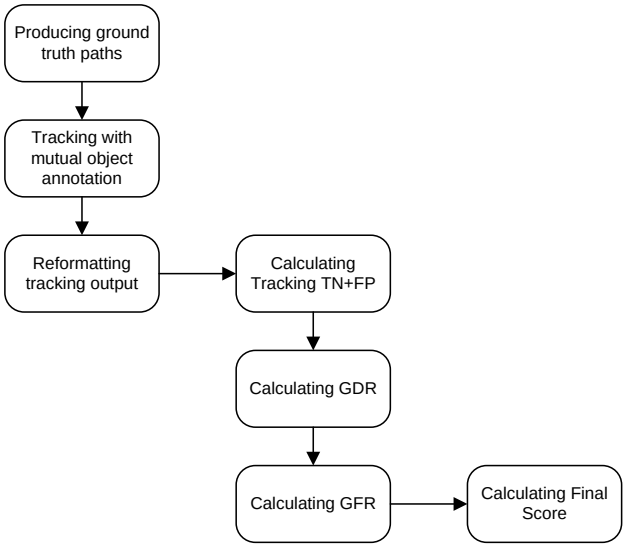


Figure 3-7 Workflow of the tracking efficiency estimation

Table 3-1 Sample of trajectory table

Path#	Frame#	MC_X	MC_Y
0	0	253.2	278.9
0	1	260.9	283.4
0	2	259.9	288.8
0	3	255.3	287.9
0	4	250.5	285.3
0	5	246.8	285.3

We illustrate the tracking efficiency of the algorithms against two types of image modalities:

1. HT/HC screen of random cell migration using epiflorescence microscopy
2. HC analysis of matrix adhesion (MA) dynamics using TRIF microscopy

The manual segmentation and tracking of each object or observation, either cell or MA, are both accomplished by biologists (observers). Image quality was first enhanced (cf. Figure 3-6b) before being processed by each observer. During the observation, each observer starts with one cell and continuously draws the outline of the same cell along the entire sequence (cf. Figure 3-6c). When cell division occurs, the observer will track both daughter cells, separately, starting from the mother cell. In that particular case, there will be two manually tracked cells that have a partial identical trajectory. In the next section, the definition of the tracking efficiency metrics will be addressed.

3.2.1. Tracking Efficiency Metrics

True Positive Tracking (TP) and False Positive Tracking (FP)

First, the true positive tracking and false positive tracking are defined for the tracking results. When comparing the tracked path to the ground truth path (cf. Figure 3-9), there are potentially three states for each time point in the trajectory (cf. Figure 3-8). The 'match' state is the good scenario at one time point in which objects from two paths are identical. The 'mismatch' state is the state at one time point in which objects from two paths are not identical. The 'no match' state is the state at one time point in which there is no corresponding object presented in either the tracked path or the ground truth path. From these three states, we define the true positive and true negative of the tracking as following:

1. The true positive tracking (cf. Equation 3-9) is measured by the number of objects in a trajectory with a 'match' state C_{match} divided by the length of the ground truth path L_{gt}.
2. The false positive tracking (cf. Equation 3-10) is measured by the number of objects in a trajectory with a 'no match' state C_{no_match} divided by the length of the ground truth path plus the length of the path being tracked $L_{gt} + L_{tp}$.
3. The true negative (TN) is calculated by $TN = 1 - FP$.

$$TP = \frac{C_{match}}{L_{gt}} \qquad \textbf{Equation 3-9}$$

$$FP = \frac{C_{no\_match}}{L_{gt} + L_{tp}} \qquad \textbf{Equation 3-10}$$

Global Detection Rate (GDR)

The global detection rate (cf. Equation 3-11) in tracking is a quantification of the success rate that each tracking algorithm can recognize all available paths. It is an indicator of the sensitivity of tracking algorithm to detect potential paths. A good tracking algorithm should not only fully recover a certain path but also recover all possible paths. Given ground truth path set T_{gt} and tracked path T_{tp}, and intersection T_{inter} between T_{gt} and T_{tp}, the global detection rate is calculated from the size of T_{inter} divided by the size of T_{gt}. The intersection T_{inter} the path overlapping calculated from tracking *TP + FP*.

$$GDR = \frac{T_{gt} \cap T_{tp}}{T_{gt}} \qquad \textbf{Equation 3-11}$$

Global Fragmentation Rate (GFR)

The global fragmentation rate in the tracking is the percentage of paths that are only tracked partially or available as fragmented trajectories. It is a supplementary quantification to global detection rate. In contrast to the global detection rate, which is an indicator of overall tracking sensitivity, the fragmentation rate is an indicator of overall tracking specificity. Figure 3-8 illustrates a partially tracked result in which a tracking algorithm produces two path fragments that should belong to one ground truth path.

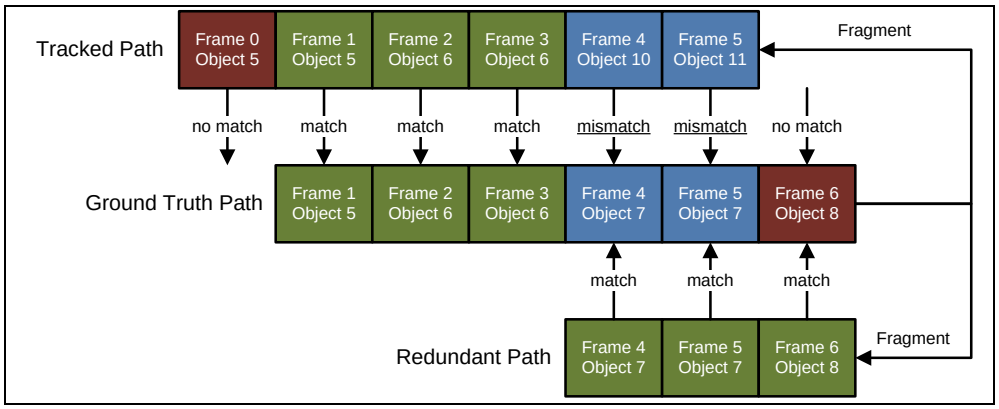


Figure 3-8 three states for each time point in one trajectory, red block: no match, green block: match, blue block: mismatch.

Final Score (FS)

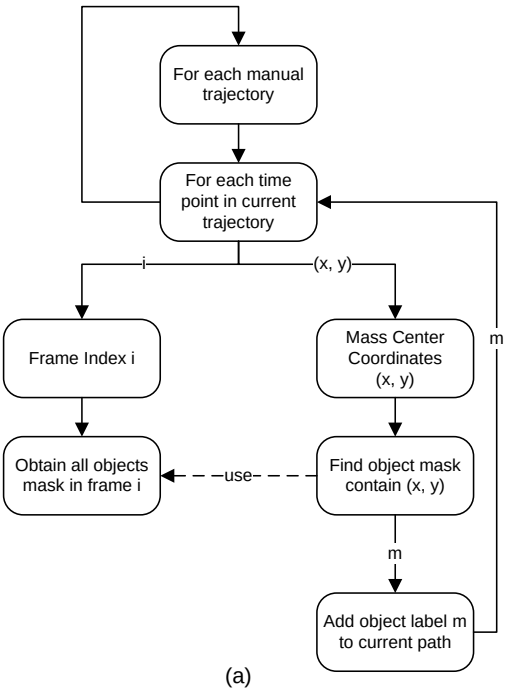
The final score of tracking is the tracking efficiency weighted by the detection rate and the fragmentation rate. It indicates the overall performance of each tracking. The F1-score [87] is calculated given tracking TP and FP (cf. Equation 3-12, Equation 3-13 and Equation 3-14) and weighted using both GDR and GFR (cf. Equation 3-15). Contrary to weighted average, the production based final score enforces a balance amongst all subsystems. The final score is equal to one only if all three subsystems are approaching to one. Unlike a weighted average, if one of the subsystems approaches zero, the whole final score will approach zero.

$$sensitivity = \frac{TP}{TP + FN} \qquad \textbf{Equation 3-12}$$

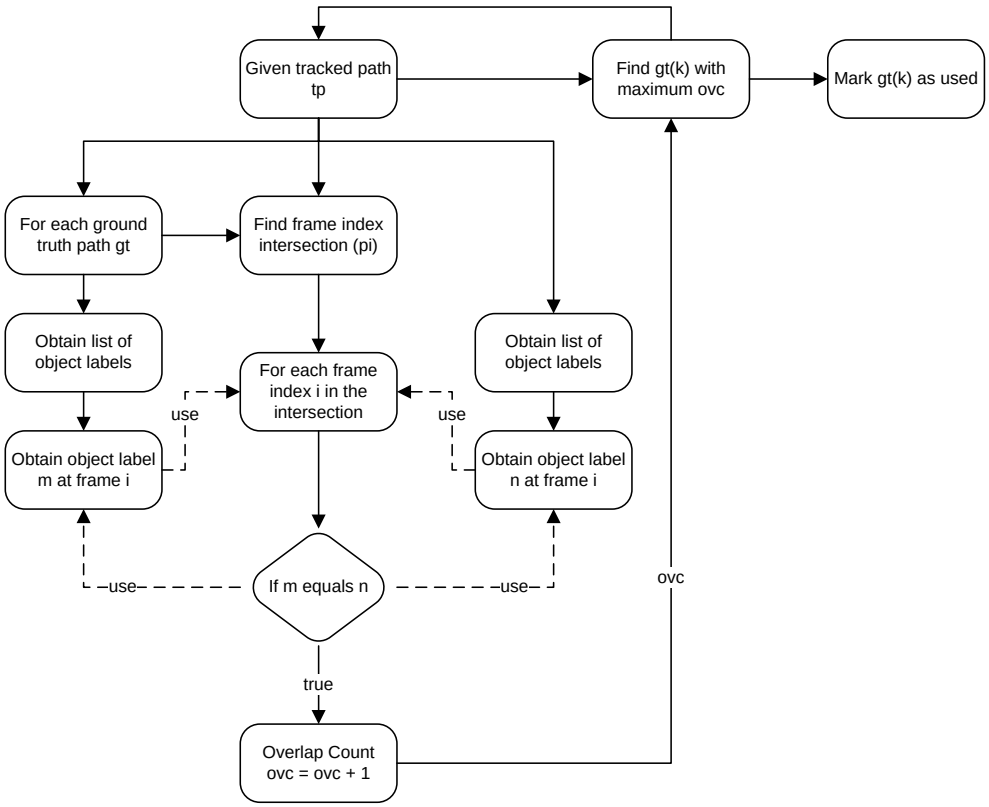
$$specificity = \frac{TN}{TN + FP} \qquad \textbf{Equation 3-13}$$

$$F1 = 2 \cdot \frac{sensitivity \cdot specificity}{sensitivity + specificity}$$ **Equation 3-14**

$$FS = GDR \cdot (1 - GFR) \cdot F1$$ **Equation 3-15**



(a)



(b)

Figure 3-9 (a) extraction of an object mask given one trajectory (b) calculating trajectory overlap

3.2.2. Results of Efficiency Estimation

MTLn3 plGFP Ground Truth

The MTLn3 plGFP is an aggressive breast cancer cell line that is regularly used in studies on migration suppression [15] and growth factor inhibition [114]. This cell line has a motile behavior and being able to track these cells is often considered a major challenge in cell biology. For the experiment, the image acquisition is accomplished using Nikon TE2000 epifluorescence microscope with a 40x lens (NA 0.75) and captured at a fixed temporal resolution of 5 min/frame.

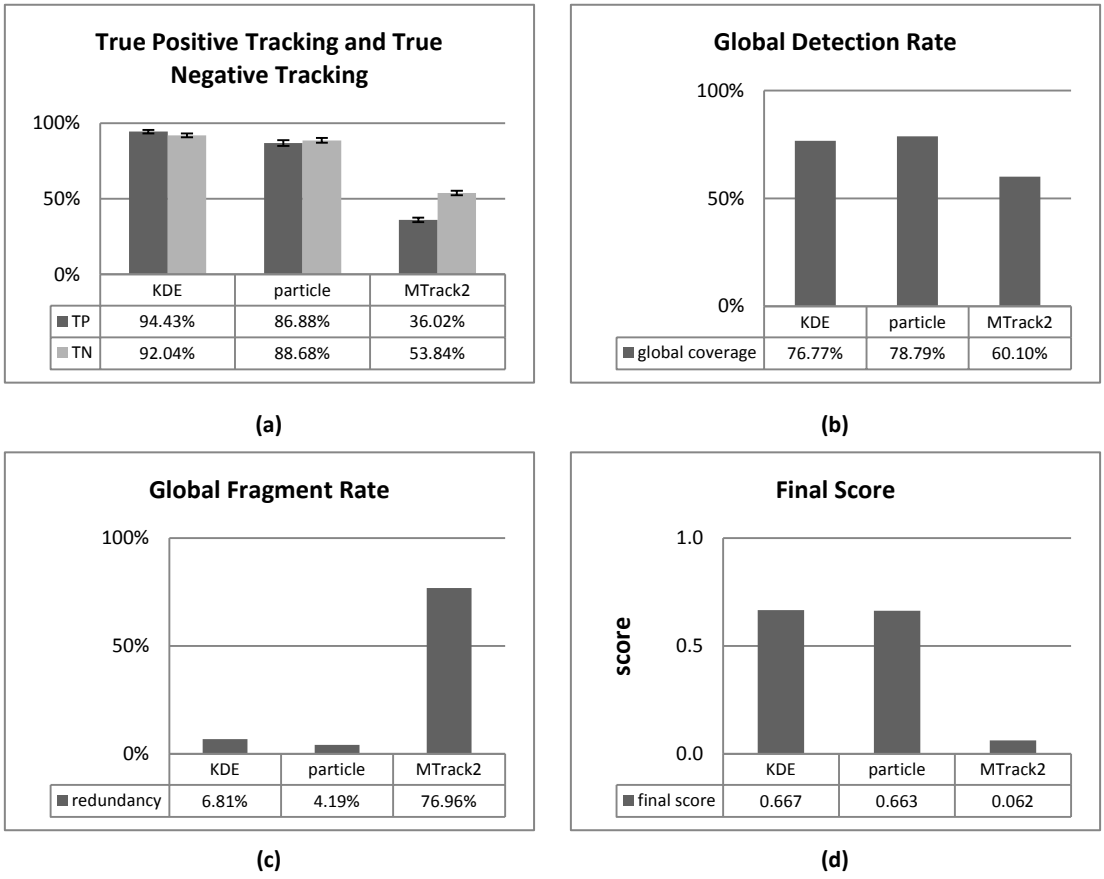


Figure 3-10 efficiency estimation using MTLn3 plGFP test set

The efficiency estimation (cf. Figure 3-10a) of each tracking algorithm shows that the KDE mean shift algorithm yields the highest TP-rate of 94.43% and the highest TN-rate of 92.04%. The particle filter tracking yields a TP-rate of 86.88% and a TN-rate of 88.68%. The MTrack2 yields the lowest TP-rate of 36.02% and the lowest TN of 53.84%. The result shows that KDE mean shift tracking produces the best overall result compared to both particle filter and MTrack2 tracking. The slightly lower performance of the particle filter tracking is mostly due to the underlying assumption of Brownian motion. The Brownian motion model is a popular function to describe random motion of particles; however, we observe that random cell migration does not necessarily follow this pattern. Therefore, particle filter tracking is less efficient compared to KDE mean shift tracking in cell migration. The performance of MTrack2 is mostly due to the variable acceleration in cell migration; a sudden acceleration of cell motility

may cause the MTrack2 to prematurely terminate a track, leading to small fragments of trajectories which, *de facto*, should be merged.

The global detection rate (cf. Figure 3-10b) shows the highest score for particle filter tracking 78.79%; KDE mean shift tracking has slightly lower score of 76.77%. The MTrack2 scores 60.10%. In our test, the results illustrate that the particle filter tracking demonstrates an acceptable sensitivity when detecting potential paths. The performance of KDE mean shift tracking algorithm is similar to the particle tracking algorithm. The MTrack2 ignores nearly half of the potential paths.

The global fragmentation rate (cf. Figure 3-10c) shows that the KDE mean shift tracking has a chance of 6.81% to produce fragmented tracking result. The particle filter tracking shows the lowest fragmentation of 4.19%. Due to its distance based criterion and cells' random acceleration, MTrack2 has the highest chance of performing a fragmented tracking.

The final score (cf. Figure 3-10d), which brings all error estimation together, shows that in general KDE mean shift tracking algorithm and particle filter tracking algorithm both demonstrate robust performance in tracking the MTLn3 plGP cell line while MTrack2 is clearly underperformed due to a prefixed tracking criterion. Generally, the performance of KDE mean shift tracking algorithm is slightly higher than the particle filter tracking algorithm. Compared to the particle filter tracking algorithm, the KDE mean shift tracking algorithm produces a result with significantly higher TP value and TN value. Since most phenotypical quantifications at the single-cell level are derived per-trajectory instead of whole population, a higher per-trajectory error may lead into wrong quantification while a redundant tracking may only shift the data distribution. Thus, it is preferable to use a tracking algorithm with high per-trajectory tracking accuracy.

MTLn3 plGFP + EGF Ground Truth

The ground truth set contains epidermal growth factor stimulated MTLn3 plGFP cells that demonstrate a significantly higher motility even compared to MTLn3 plGFP control cells. The cells are moving approximately five-times faster than MTLn3 plGFP control cells and twenty-times compared to other cell lines. The current image acquisition is accomplished using the Nikon TE2000 epifluorescence mode with a 40x lens (NA 0.75) and captured at a fixed temporal resolution of 5 min/frame. The relatively high motility requires a higher temporal resolution than the actual resolution used. Consequently, manual segmentation and tracking of MTLn3 plGFP +EGF is considerably more difficult.

The tracking error (cf. Figure 3-11a) shows that the tracking efficiency of both the KDE mean shift and the particle filter tracking algorithm is significantly decreased while the performance of MTrack2 is significantly increased. The major reason is due to that cells may move out of the field of view; then the MTrack2 will terminate the tracking while the other two algorithms may attempt to resume the tracking procedure. This also leads to the lower global detection rate of KDE mean shift tracking algorithm (cf. Figure 3-11b) since KDE mean shift intends to combine two ground truth paths into one tracked path when one ground truth path escapes the image frame. Moreover, although the early termination mechanism of MTrack2 may improve

tracking accuracy in MTLn3 plGFP +EGF, it receives more penalties from the GFR (cf. Figure 3-11c). The final score of MTLn3 plGFP +EGF (cf. Figure 3-11d) shows a similar tendency compared to the results of MTLn3 plGFP.

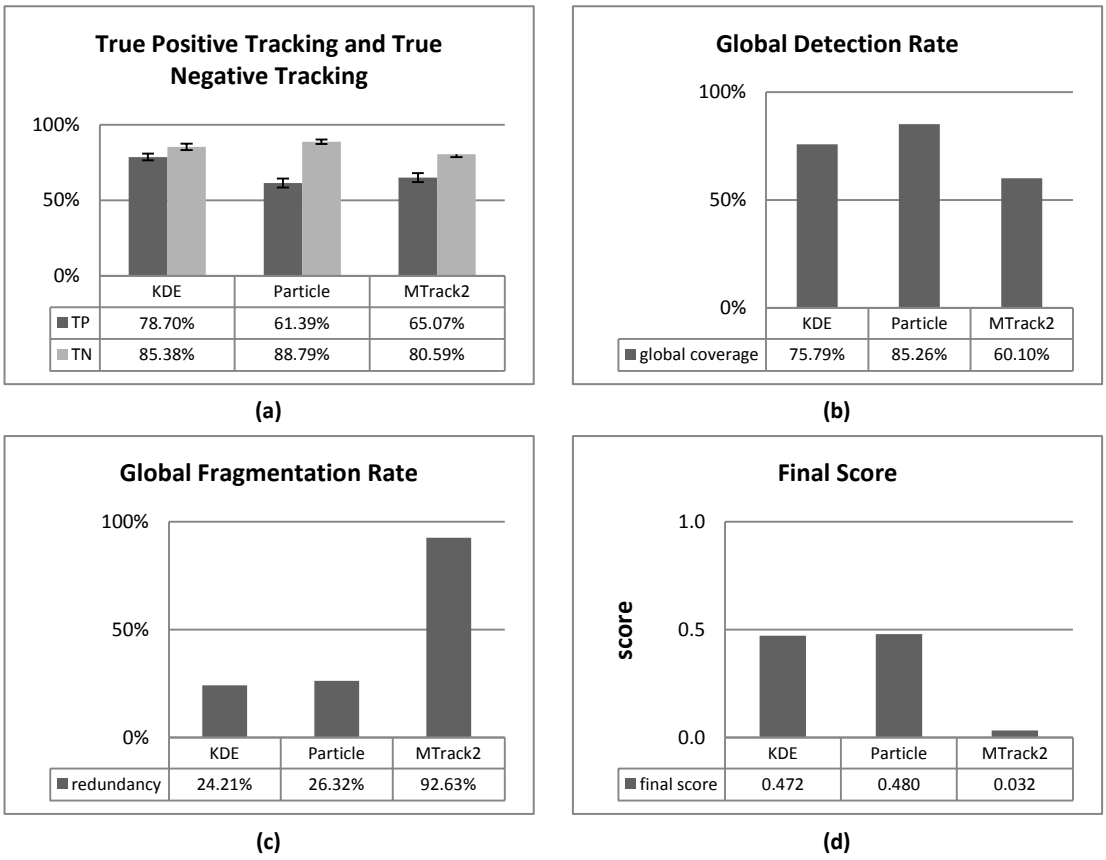


Figure 3-11 efficient estimation using MTLn3 plGFP+EGF test set

It may be that the KDE mean shift tracking provides a more accurate per-trajectory tracking. However, as the method lacks an early-termination mechanism and escaping detection in the implementation used for testing, its global performance is slightly lower than particle filter tracking. The particle filter tracking algorithm produces a result containing considerable amount of per-trajectory errors. In HT/HC study, it is preferred to have a low per-trajectory error over a low global error since most phenotypical quantifications are derived from each trajectory and not from the whole population. The MTrack2 produces low quality results due to highly aggressive cell behavior and non-linear shape deformation, and resulting in the least desirable outcome. However, from time-lapse sequence with better temporal resolution and less motile cells, it is still a computation inexpensive solution. Moreover, it is foreseeable that with both early termination mechanism and escaping detection implemented, the performance of KDE mean shift tracking algorithm should be higher than both particle filter tracing and MTrack2.

Matrix Adhesion Ground Truth

This ground truth set contains dynamic matrix adhesion (MA) [16][36]. The analysis of MA dynamics is an essential part of cell migration study (cf. Ch. 5). Compared to cell tracking, MAs have a simpler and monotonous morphology. There is only a limited amount of information

that can be used to distinguish two MAs. It complicates the tracking procedure since morphological heterogeneity between potential candidates is reduced, therefore leading to a higher false positive rate when building object linkages.

Based on per-trajectory error estimation (cf. Figure 3-12a), it shows that energy-driven linear tracking algorithm has lower per-trajectory error compared to particle filter and MTrack2. In general, all three tracking algorithms can produce acceptable per-trajectory tracking accuracies. The slightly lower performance of particle filter tracking is mainly due to the assumption of Brownian motion model whereas of a linear motion model better fits the nature of MA dynamics [36].

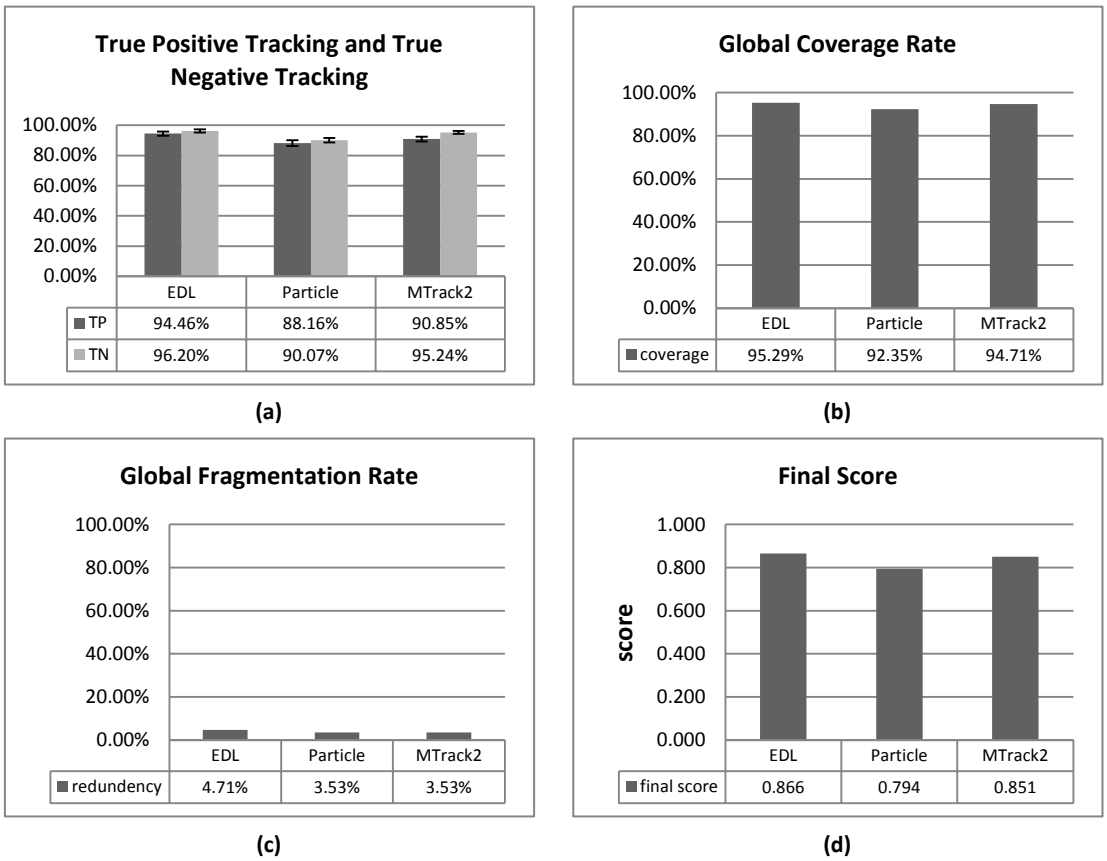


Figure 3-12 efficient estimation using matrix adhesion test set

The global coverage rate (cf. Figure 3-12b) shows that the energy-driven linear tracking algorithm has the highest global coverage of 95.29%. The difference between the three algorithms, however are overall the same.

The global fragmentation rate (cf. Figure 3-12c) shows that the energy-driven linear tracking has a result of lower quality whereas the particle filter and MTrack2 are slightly better. The result suggests that tracking result of EDL may contain more early-terminated trajectories compared to the particle filter and MTrack2 tracking.

The overall performance (cf. Figure 3-12d) shows that the EDL has the best score compared to the particle filter and the MTrack2 tracking algorithm. However, the performance of MTrack2

shows similar performance as EDL. We conclude that both EDL and MTrack2 use all available information from binary mask and intensity landscape while particle tracking must rely on Brownian motion model which seems to be a false assumption in this particular tracking problem.

3.2.3. Temporal Resolution Variance

Temporal resolution is a measurement of video sampling rate. The ratio between temporal resolution and object velocity plays a crucial role in video tracking. A higher velocity-to-temporal-resolution ratio (VTR) suggests that intermediate stage of object motion may be overlooked. As a result, a higher VTR means more variations and fewer similarities between objects in consecutive frames. Here we propose the **Consecutive Displacement Rate (CDR)** measurement of the VTR. The CDR is considered as a numeric indicator of tracking complexity.

Consecutive Displacement Rate

The consecutive displacement rate (CDR) is the pixel-based intensity-weighted difference between consecutive objects. Given consecutive objects A and object B, the CDR is expressed as:

$$CDR = \frac{\overline{A \cap B}}{A \cup B}$$ **Equation 3-16**

From the algorithm descriptions (cf. §3.1) we can conclude that all tracking algorithms rely on certain similarity criteria when constructing a linkage between consecutive objects. A high CDR suggests that two objects share less overlap, thus, it suggests a potential shape change or significant position shifting. A significant position shifting may introduce an error to the tracking algorithm that relies on relative position such as blob tracking and particle filter tracking while a potential shape change causes a problem to the algorithm such as KDE mean shift tracking. Therefore, a good time-lapse image data set should express the lowest CDR as possible.

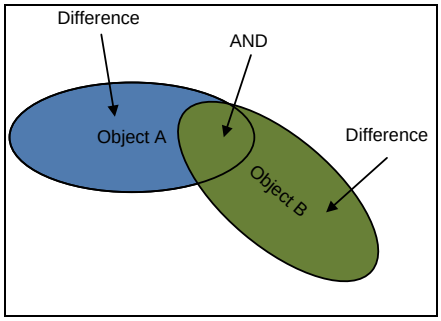
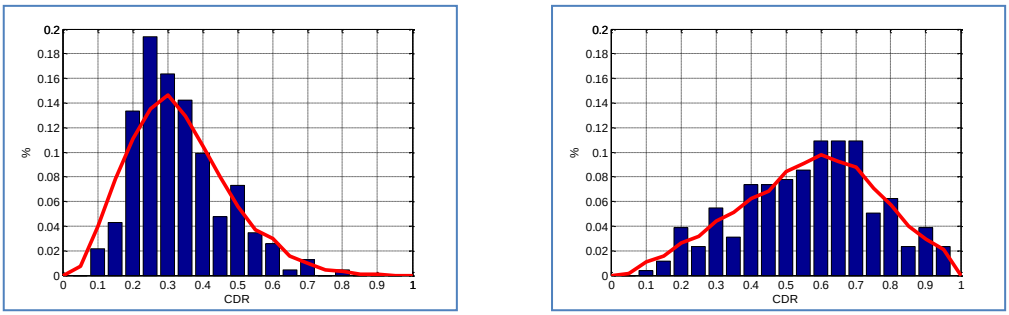


Figure 3-13 object difference and mutual area



(a) MTLn3 plGFP ctrl **(b) MTLn3 plGFP EGF**

Figure 3-14 CDR of MTLn3 plGFP cell line under different conditions, red lines are tend lines by moving average

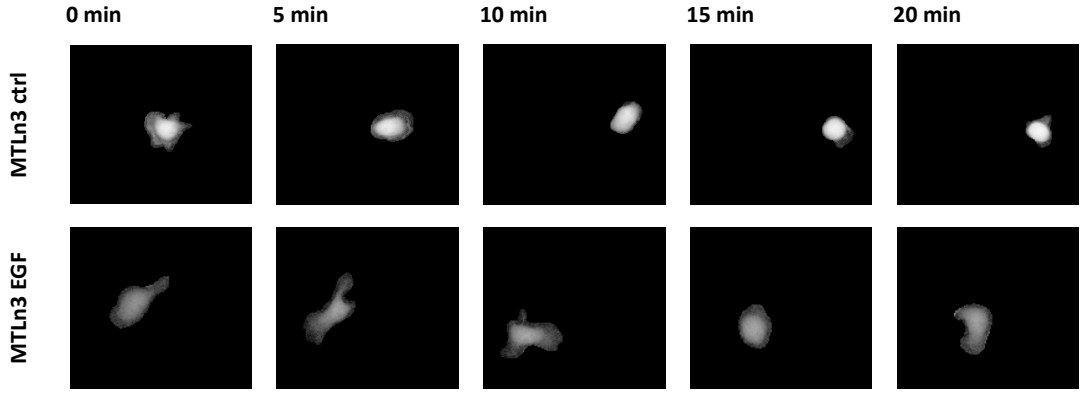


Figure 3-15 snapshots of time-lapse image sequence of MTLn3 GFP channel

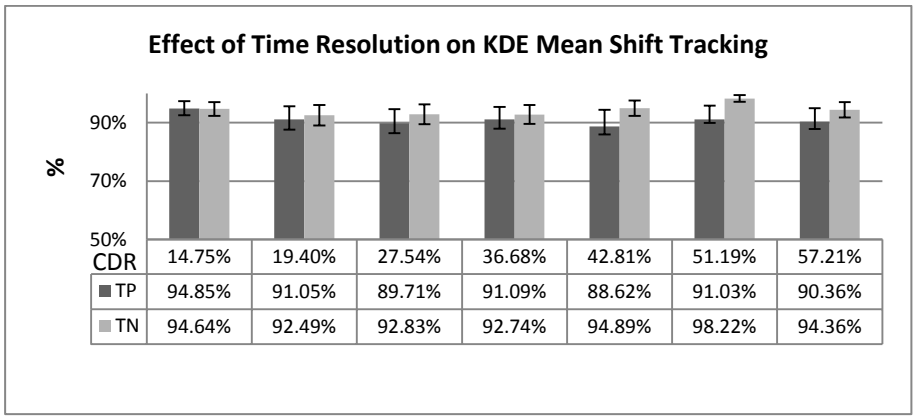
Figure 3-14 shows that the CDR distribution of EGF stimulated cells significantly shifts to the right side of CDR distribution of control cells, suggesting at current temporal resolution the tracking of EGF stimulated cells becomes less efficient. It also shows that between consecutive time points EGF stimulated cells have more than 60% variation in either position or shape. Moreover, snapshots of the MTLn3 set (cf. Figure 3-15) clearly show that migration behavior of EGF stimulated cell is more aggressive and therefore difficult to model.

Subsampling Performance Estimation

To further characterize the effect of temporal resolution on tracking efficiency, a control test has been designed. The control test consists of the following steps:

1. Capture image sequence with high temporal resolution (30 second/frame)
2. Create subsampled image sequence by extracting one frame every r^{th} frame
3. Estimate CDR of subsampled image sequence (cf. Equation 3-16)
4. Perform tracking on the subsampled image sequence
5. Compare tracked path to ground truth path (cf. § 3.2.1)

The effect estimation suggests that tracking true positive slightly decreases at increasing CDR (cf. Figure 3-16). This also suggests that tracking efficiency will be affected when complexity of object motility increases. We consider KDE algorithm for tracking with different CDR. From the test result, we observe that even with a relatively high CDR, the KDE algorithm still produce a tracking result of TP = 90.36% and TN = 94.36% (cf. Figure 3-16).



CDR	14.75%	19.40%	27.54%	36.68%	42.81%	51.19%	57.21%
TP	94.85%	91.05%	89.71%	91.09%	88.62%	91.03%	90.36%
TN	94.64%	92.49%	92.83%	92.74%	94.89%	98.22%	94.36%

Figure 3-16 effect of time resolution on KDE mean shift tracking efficiency

3.3. Conclusions and Discussion

In this chapter, we have discussed four tracking algorithms that can be employed in the object tracking of HT/HC screens. The performance assessment of tracking algorithms shows that all tracking algorithms can produce acceptable tracking results for less motile objects such as MTLn3 plGFP cells. Our dedicated solutions, namely the KDE mean shift tracking algorithm and energy driven linear tracking can still produce a near optimal overall performance in aggressive cell behavior in an MTLn3 plGFP +EGF image set compared to other existing solutions.

Further analysis of assessment results shows that irregular and local object deformation poses difficulties for confidence measurement based tracking algorithms. The robustness of confidence measurement is a determining factor behind successful tracking. In this case, the KDE mean shift algorithm clearly outperforms the blob tracking algorithm since the kernel density estimation is more flexible and robust to small shape deformation; e.g. deformation by cell protrusions.

Compared to confidence measurement based tracking, motion model based tracking algorithms works best in object tracking when object shape information is limited, typically the matrix adhesion (MA) tracking. In the MA tracking scenario, confidence measurements between true and false candidates are less discriminative since the sizes of objects are around 5~30 pixels. However, motion model based tracking algorithm takes advantage of the presence of shape information in tracking aggressive cell migration. Moreover, the assumption of a Brownian motion model or linear motion model is not necessary true for cell migration. The major drawback behind the motion model based tracking is the assumption of a fixed motion model; increasing robustness of motion model based tracking will require flexibilization of the motion model.

From the advantages of each algorithm, we further conclude that the selection of a proper tracking algorithm should not only be based on a quality factor such as CDR. It should also be based on the availability of information on object shape and motion model. When information is limited, the choice of the tracking algorithm must be changed accordingly. Instead of considering either a confidence measurement or a motion model, a tracking solution combining both types of linkage criteria may produce more robust tracking results.

Acknowledgement

The authors would like to thank Ying Shi for her meticulous work on making the benchmark data

Chapter 4

A Study to Cell Migration Analysis

This chapter is based on the following publications

Le Dévédec, S.E., Yan K., de Bont H., Ghotra V., Truong H.,Danen E., Verbeek F.J., and vande Water B.(2010), "A Systems Microscopy Approach to Understand Cancer Cell Migration and Metastasis", *Cellular and Molecular in Life Science*, Vol. 67, Issue 19, pp.3219-3240

Damiano, L., Le Dévédec, S.E., Di Stefano, P., Repetto, D., Lalai, R., Truong, H., Xiong, J.L., Danen, E.H., Yan, K., Verbeek, F.J., De Luca, E., Attanasio, F., Buccione, R., Turco, E., van de Water, B. and Defilippi, P.(2011), "p140Cap suppresses the invasive properties of highly metastatic MTLn3-EGFR cells via impaired cortactin phosphorylation", *Oncogene*, Vol. 31, Issue 5, pp.624-633. doi: 10.1038/onc.2011.257.

Chapter Summary

In this chapter, we focus on estimating the practical performance of an image analysis solution within the scope of a live cell HT/HC screen study of tumor cell morphology and motility. In detail, we address the following research question:

1. Can the unique phenotypical profile of cells treated with different growth factors be characterized using newly developed algorithms for image segmentation (cf. Ch.2) and object tracking (cf. Ch. 3) in HT/HC screen?

Following the workflow of HT/HC screen (cf. Figure 1-2), this chapter is divided into three major sections. First, the experimental design and image acquisition procedure are illustrated. The selected growth factors and their expected cell responses are explained. Then the design of the image analysis solution is demonstrated and the motivation behind this design is explained. Next, the measurements of cell morphology and motility extracted through image analysis are illustrated. Finally, the variation in morphology and motility are derived from numerical measurements using data analysis. These results are compared to morphology and motility as described in the literature.

4.1. Workflow of Growth Factor Analysis

4.1.1. Experiment Design

Cell morphology and motility are critical parameters in many physiological as well as pathophysiological processes. Recent progress in the technology of live cell enables the capture of detailed information on morphology and motility for cell biology analysis. This chapter aims to use functional genomics and compound screening to unravel the mechanisms of tumor cell migration in the context of breast cancer metastasis. To that end, an image analysis solution is often used to provide an automated analysis solution for such HT/HC *in vitro* live cell screen. Central to this solution is the segmentation algorithm [62] (cf. Ch. 2). From the segmentation, a per-cell tracking over the time-lapse sequence is accomplished via tracking algorithm [9] (cf. Ch. 3). From the tracked cells, a set of numeric measurements related to tumor cell behavior is extracted.

We validated this method through a pilot experiment that assays the random migration of highly motile tumor cells with growth factor regulation. As results, we obtained a sensitive, time-resolved profiling for every condition. During invasion and metastasis, migration is driven by growth factors, such as Epidermal Growth Factor (EGF), Fibroblast Growth Factor (FGF), Hepatocy Growth Factor (HGF) and Transforming Growth Factor β1 (TGF-β1). All growth factors (GFs) are well known chemoattractants or cytokines involved in various cellular processes. However, the response of tumor cells at the level of cell migration to those growth factors, as supplemented individually or in combination, has not yet been quantified. To understand how those different GFs control the tumor cell motility, we performed an experiment that includes single exposure to the different GFs and combinations of GFs.

Cell Culture

Two subgroups of MTLn3 rat breast cancer cells [115] were cultured as previously described [116]. The MTLn3-plGFP is a standard MTLn3 cell line while the MTLn3-GB1 is an MTLn3 cell line with overexpression of the EGF-receptor. These MTLn3 cell-lines were previously described in other tumor migration studies [115][80]. They were maintained in αMEM (Life Technologies, Inc., Gaithersburg, MD) supplemented with 5% fetal bovine serum (Life Technologies). During the random cell migration assay, the cells are exposed to different growth factors known to be involved in tumor progression in general and in tumor cell motility in particular. We analyzed the motility behavior of MTLn3 cells when exposed to 12 different treatments including: DMSO, EGF, FGF, HGF, TGF-β1, EGF+FGF, EGF+HGF, EGF+ TGF-β1, FGF+HGF, FGF+ TGF-β1, HGF+ TGF-β1, EGF+FGF+HGF+TGF-β1. The growth factors used are as follows: Epidermal Growth Factor (EGF, 100 µg/ml), Fibroblast Growth Factor (FGF, 100 µg/ml), Hepatocy Growth Factor (HGF, 20 µg/ml) and Transforming Growth Factor β 1 (TGF-β1, 20 µg/ml). The overall plate design is given in Table 4-1. The DMSO treatment is considered the negative control group.

Table 4-1 Plate design of GF regulation experiment (plGFP-yellow, GB1-green), alphabet represents column index, number represents row index

	plGFP				GB1			
	C	D	E	F	G	H	I	J
2	DMSO	HGF	EGF+HGF	FGF+TGFβ	DMSO	HGF	EGF+HGF	FGF+TGFβ
3	DMSO	HGF	EGF+HGF	FGF+TGFβ	DMSO	HGF	EGF+HGF	FGF+TGFβ
4	EGF	TGFβ	EGF+TGFβ	HGF+TGFβ	EGF	TGFβ	EGF+TGFβ	HGF+TGFβ
5	EGF	TGFβ	EGF+TGFβ	HGF+TGFβ	EGF	TGFβ	EGF+TGFβ	HGF+TGFβ
6	FGF	EGF+FGF	FGF+HGF	all	FGF	EGF+FGF	FGF+HGF	all
7	FGF	EGF+FGF	FGF+HGF	all	FGF	EGF+FGF	FGF+HGF	all

Live Cell Imaging

Glass bottom 96-well plates (PAA) were coated with 20 µg/µl collagen type 1 (isolated from rat tails) for 1 hr at 37 ºC. Cells were plated directly onto the coated glass coverslips and imaged 24 hrs after plating. Prior to the imaging, the medium was refreshed using serum free medium to starve the cells for at least 4 hrs. Cells were then placed on a NIKON Eclipse TE2000-E widefield microscope using a 20x objective lens (0.75 NA, 1.00 WD), perfect focus system, and a 37 ^{o}C incubation chamber. These cells are directly exposed to the different growth factors at the beginning of the experiment. The image acquisition was done with 3 locations per well and GFP signal was acquired every 5 min for a total imaging period of 12 hrs. The images are captured using a NIKON DQC-FS EMCCD camera with 16-bit pixel depth in 512x512 pixels. For each plate, we collected 144 time-lapse image sequences and this experiment was triplicated.

4.1.2. Image Analysis

For each plate, 144 time-lapse sequences are acquired. With a total number of three replicas, the MTLn3 Growth Factor Regulation experiment produced over 432 image sequences, each containing 145 frames. The whole experiment produced over 50k images. Therefore, the use of automated image analysis is essential. The major goal of image analysis solution is to automatically extract both motility and morphology quantifications from these images. To that end, we need to both extract individual cells and track them using an integrated solution combining both image segmentation and object tracking. These HT/HC images are first uploaded to the image management server for the sake of efficient data management. The pipeline is illustrated as four consecutive steps: (1) **image preprocessing**, (2) **image segmentation**, (3) **object labeling**, and (4) **object tracking**.

4.1.2.1. Image Preprocessing

For each image sequence, an image preprocessing procedure is first applied to improve image quality for segmentation using noise suppression or signal enhancement algorithms. The image preprocessing procedure contains four sequential steps (1) **intensity normalization**, (2) **median filter**, (3) **Gaussian filter,** and (4) **Rolling ball filter**. Each step removes or reduces image quality issues during the image acquisition. In the images (cf. Figure 4-1a), we observe an uneven illumination. In the cells, we observe an intensity variation. In addition, a high level Poisson noise [28][57][109] is observed in the images.

Intensity normalization (cf. Figure 4-1b) is first employed to rescale the intensity values over the complete dynamic range. Intensity normalization is used to ensure that the entire dynamic

range is used by all images. The most common intensity normalization formula is described as following:

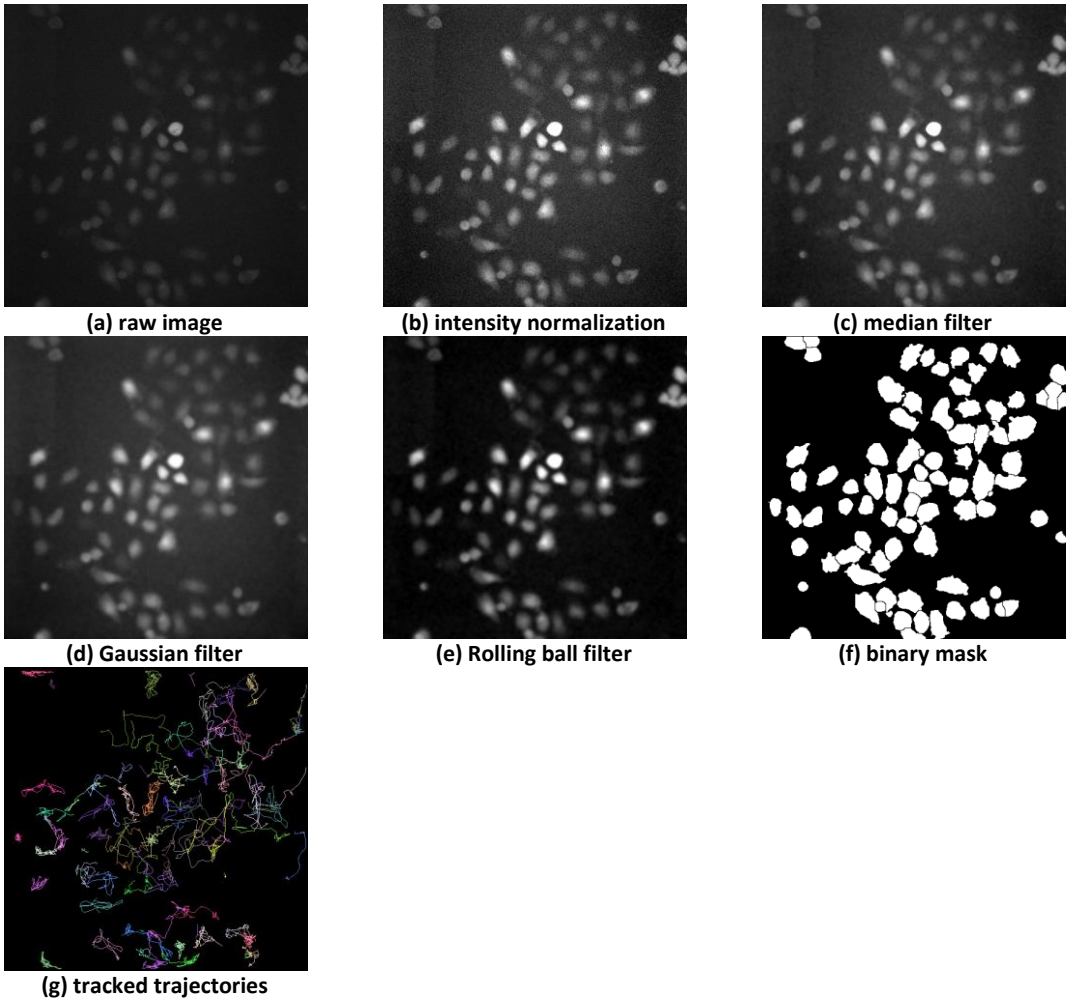
$$I' = \big(I - min(I)\big) \cdot \frac{2^n - 1}{max(I) - min(I)}$$ **Equation 4-1**

, where I is the intensity values in the image and n is the bit-depth of the image.

After the intensity normalization, a **median filter** [85] is applied to the image (cf. Figure 4-1c). A 3x3 median filter kernel [6] can effectively removes the Poisson noise [28][57][109].

Subsequently, a **Gaussian filter** is applied ($\sigma = 2$, just above the Poisson noise) to the image (cf. Figure 4-1d) to create a smoother and more continuous intensity landscape for the seeded watershed algorithm.

Finally, a **rolling ball filter** is applied to the image (cf. Figure 4-1e) to remove the uneven illumination due to autofluorescence [117][101] (cf. Figure 4-1b) from the culture medium. The current rolling ball filter uses a spherical kernel with a radius of 128 pixels so that the kernel is approximately little larger than a cell, leaving the dominant foreground signal. After preprocessing, the images are ready for segmentation.



(a) raw image **(b) intensity normalization** **(c) median filter**

(d) Gaussian filter **(e) Rolling ball filter** **(f) binary mask**

(g) tracked trajectories

Figure 4-1 intermediate results of image analysis

4.1.2.2. Image Segmentation

An image quality assessment with 32 randomly selected images from the experiment (cf. Figure 4-2 raw images) shows that:

1. Regardless the image preprocessing, object intensities are not consistent within the same image sequence. The object intensities are not consistent even for an object at different time points. Such variability seems to be related to different level of GFP expression in the cell [118][21].

As a result, this image set has an average **coefficient of variation (CV) of 4.3** (cf. Ch. 2 Equation 2-8); consequently it is difficult to ensure a high score for both sensitivity and specificity. Therefore, the dedicated solution **WMC segmentation algorithm** (cf. Ch. 2) is used since it is designed to handle intensity variation. From the WMC algorithm (cf. Figure 4-1f), each object is labeled[119][120]. All labeled objects will be passed to the object tracking solution.

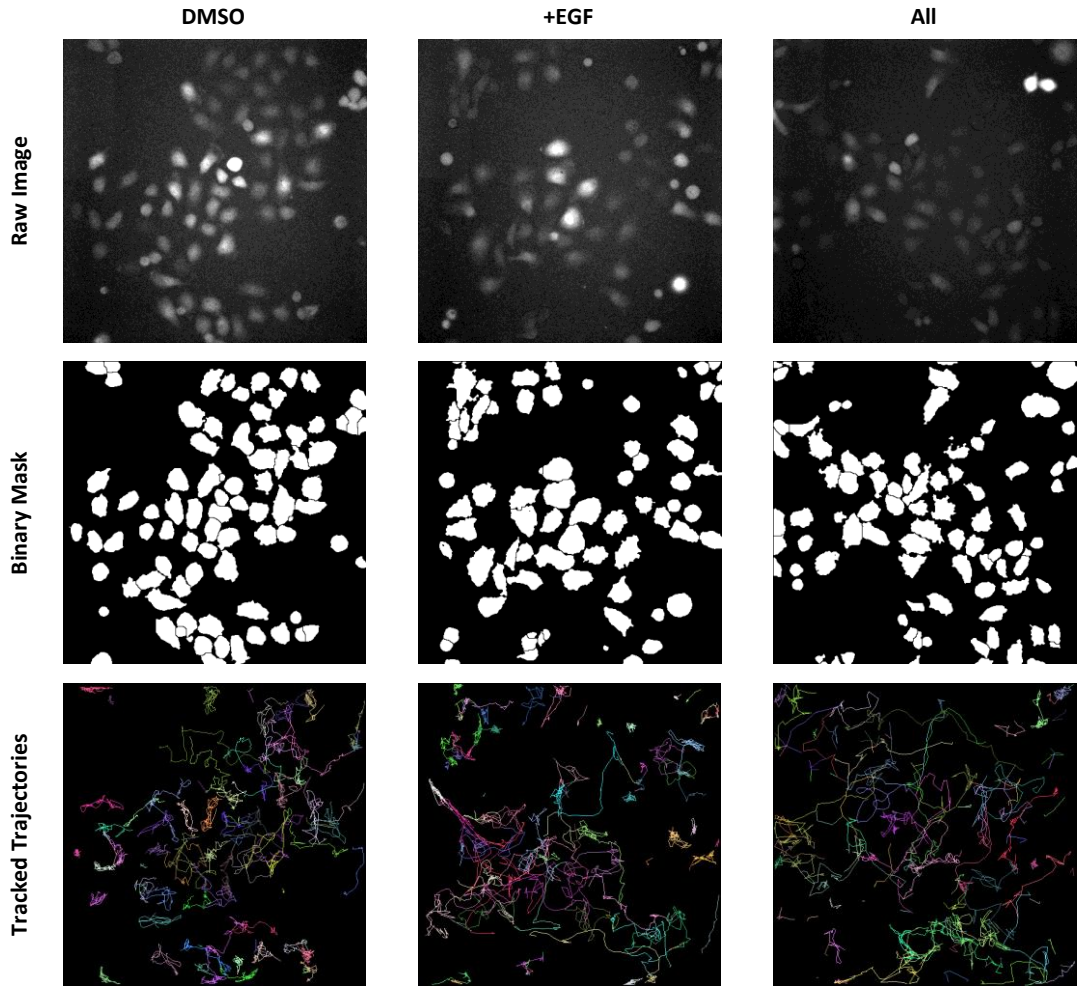


Figure 4-2 segmentation and tracking results for the different GF treatments

4.1.2.3. Object tracking

Cell tracking with this image set faces several limitations. First, since this experiment is a study of cancer cell migration behavior under the influence of a growth factor, the motility of these cells (1.2~4.2 μm/min) [9][10] is higher compared to existing studies [121] (0.1~0.4 μm/min). Confidence measurement based tracking solutions such as overlap tracking cannot be easily adapted since the confidence measurement no longer applies. The variation in the cell motility

under different treatments complicates the performance of motion model based tracking. Moreover, during cell migration a cell is exposed to non-rigid deformations. As a result, the image set has an average **consecutive displacement rate (CDR) of 0.3** (cf. Ch. 3 Equation 3-16), which for the temporal-resolution is barely sufficient to capture the continuity of cell body deformation and position shift during cell migration. Therefore, here we employ the robust tracking solution **KDE mean shift tracking algorithm**. The KDE algorithm can produce a higher overall performance and robustness in tracking cells with constantly changing migration patterns (cf. Figure 4-2 tracked trajectories). Once consecutive objects are tracked, measurements will be further extracted from both binary masks and trajectories.

4.1.2.4. Morphology and Motility Measurement

For image analysis two categories of measurements are introduced: the morphology measurements and motility measurements [9][10]. The morphology measurements are quantifications describing the shape change of each object. They are derived from the binary mask and the associated intensity information of the original image. The motility measurements are quantifications describing the object motion pattern. They are derived from trajectories obtained from the object tracking algorithm. With both motility and morphology measurements, we will further quantitatively compare cell behavior under different treatments (cf. Figure 4-2).

4.1.3. Data Analysis

We designed a data analysis pipeline (cf. Figure 4-3) to extract comprehensive information from the data. The data analysis pipeline converts morphology and motility measurements into a representation to support the research question. The data analysis pipeline consists of: (1) overview, (2) target identification, (3) individual comparison, and (4) manual verification.

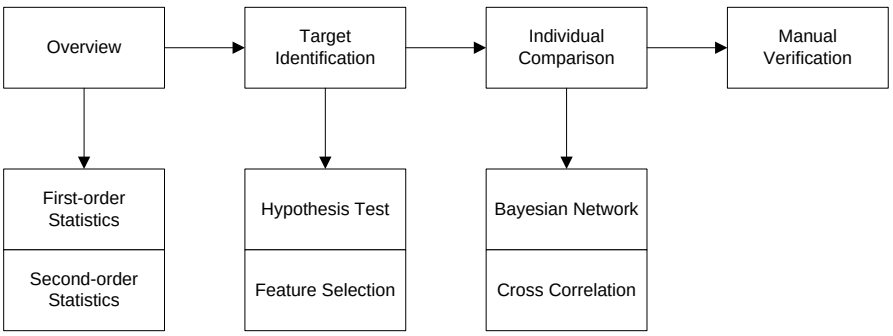


Figure 4-3 data analysis workflow of cell migrations

Overview

In the overview, HT/HC measurements from different conditions and locations are combined and visualized to provide an overall pattern for the whole data set. The combination of the data set in the overview step is accomplished using low-order statistics such as mean or median, and the second-order statistics such as variance or correlation [14][102][122][12][123]. In practice, they are considered an efficient and fast representation over a large amount of data [10][15][124]. As a generalization of cell behavior, it is still rough.

Target Identification

In the target the identification step, measurements are coupled for statistical comparison in order to verify whether these measurements are different by chance. Statistical significance tests [124][125][126] [127] or feature selection solutions (cf. Figure 4-4) [10][15][33][12] are used in this step.

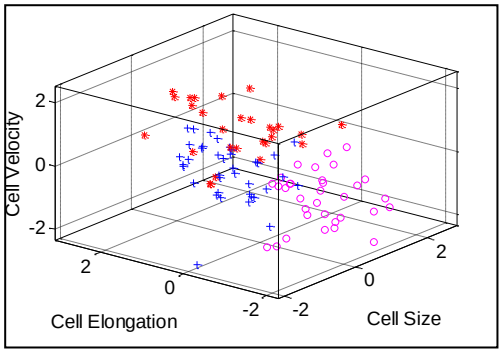


Figure 4-4 top-3 measurements selected by feature selection between ctrl (blue), EGF treated (red), and HGF treated (magenta) cells.

Temporal Profiling

The low-order statistics do not reveal the temporal tendencies in the measurements. Therefore, temporal-order statistics [39][12] are further used as dynamic modeling solution for cell behavior. The typical solutions of cell behavior, i.e. a Dynamic Bayesian Network (DBN)[128][129][130] or a Hidden Markov model (HMM) [33][12], are frequently used in the modeling of spatio-temporal patterns [33][111].

User Verification

The manual verification is the final step in data analysis. The researchers perform a check of the image analysis results of the interesting targets. They also perform a similar check over the control condition. In HT/HC, a behavior that is significantly altered may suggest an interesting target. User verification over these targets will provide further understanding of the data. It is possible that the image and data analysis may result in errors due to experimental bias or design flaws. The user verification of image analysis in both control condition and treated is considered an obvious necessity. We currently use the following checklist during user verification:

- Too few foreground pixels being captured?
- Too many background pixels being captured as foreground?
- Most individual objects separated correctly?
- Most objects captured?
- Most object trajectories tracked?
- Too many trajectories overlapping?
- Too many early terminations of trajectories?
- Is change/difference of measurement corresponding to observation for controls?

In our approach, researchers are presented with a number of overlay images from each condition (cf. Figure 4-5) enabling a quick overview on the performance of segmentation and tracking frame-by-frame. In practice, users check the segmentation result by looking to how the overlay mask fits the user perception of the image content (cells) (cf. Figure 4-5 mask overlay). The trajectories are further overlaid (cf. Figure 4-5b) to the mask-overlay image (cf. Figure 4-5a) to allow users to check the tracking performance.

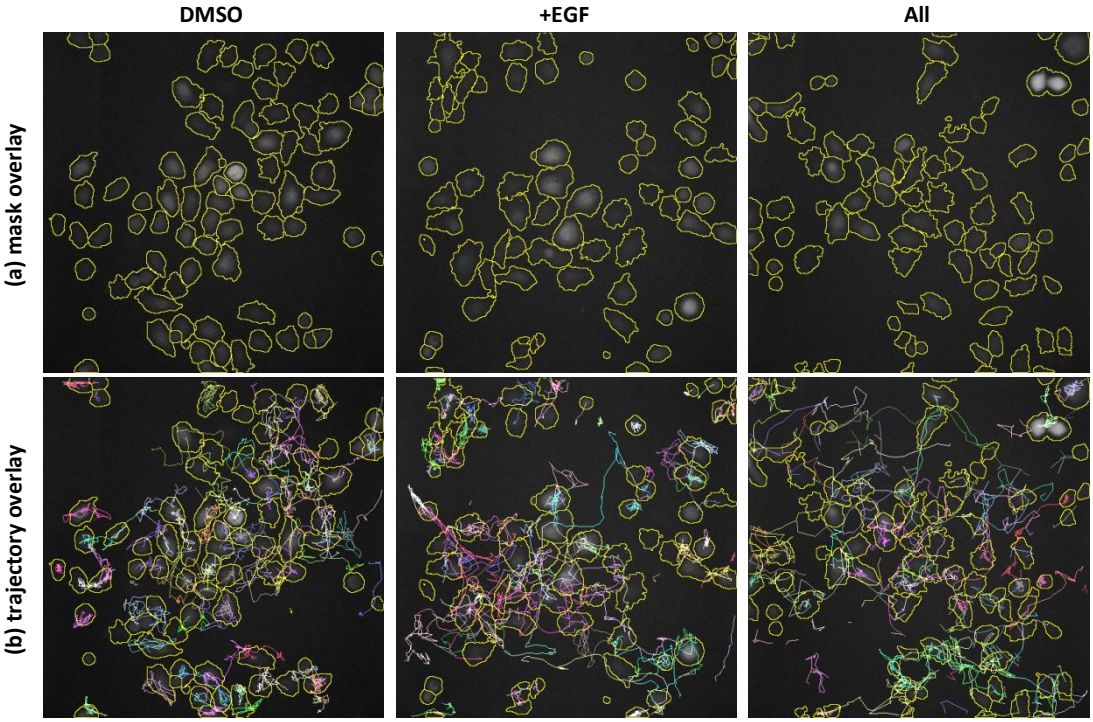


Figure 4-5 manual verification of segmentation and tracking results of the targets

4.2. Analysis of GF Regulation

4.2.1. Differential effect of individual and combined GF stimulation

In Figure 4-6, a heat map is generated from the mean values of all features and a hierarchical clustering is performed. From the heat map, it is clear that HGF treatment as a single growth factor or in combination with other GFs is the most potent regulator of the migration of MTLn3 cells. Indeed, HGF alone and other combinations are all clustered together and the clustering is based on an increase in velocity and motion linearity (cf. §4.2.2 persistence). While EGF is not directly inducing an increase in cell velocity, it seems to affect the directionality of the cells. Surprisingly, TGF- β1 does not seem to have any measurable effect on the migration of MTLn3 cells. Finally, FGF does affect cell size accompanied with a slight effect on motion linearity (cf. Figure 4-2).

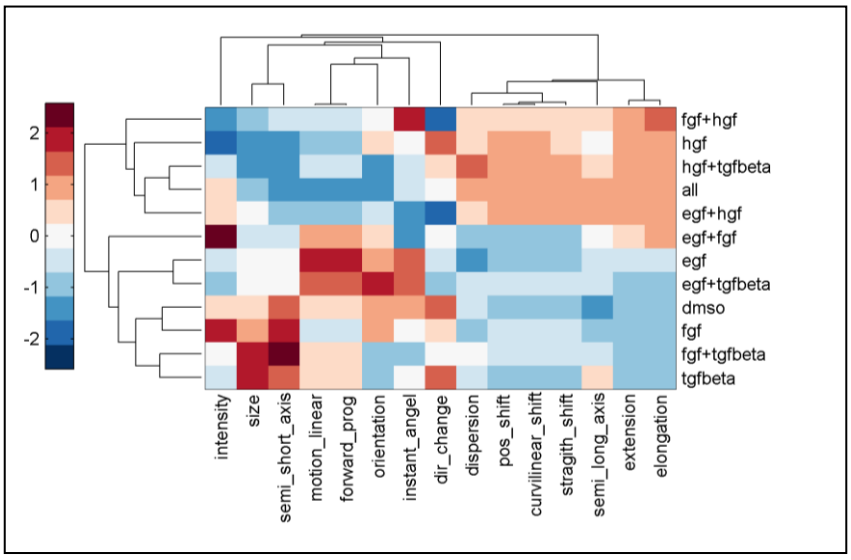
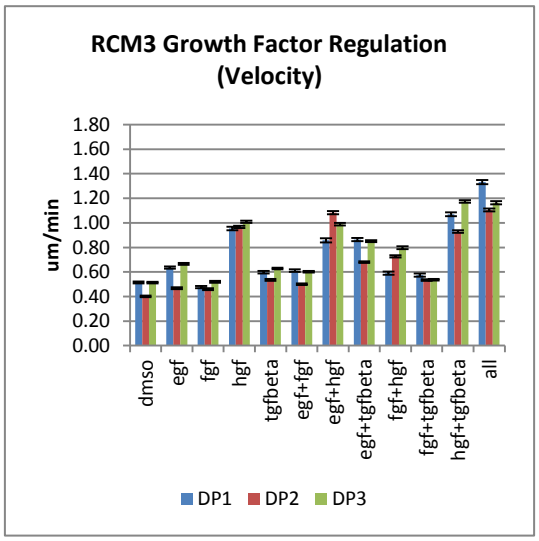


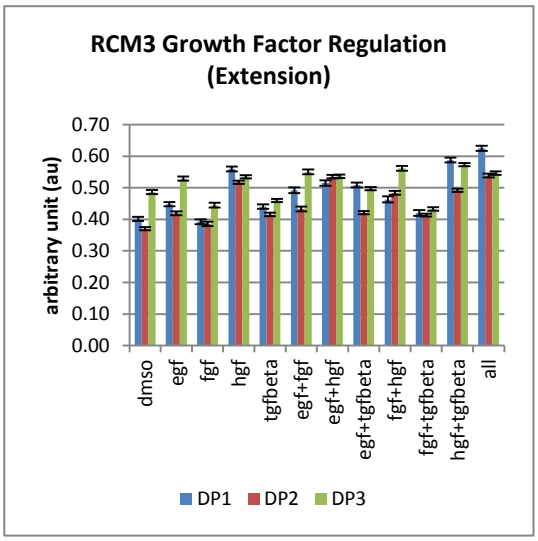
Figure 4-6 heatmap of the RCM3 data plGFP

4.2.2. HGF stimulation increases speed while EGF increases the migration persistence of MTLn3 cells

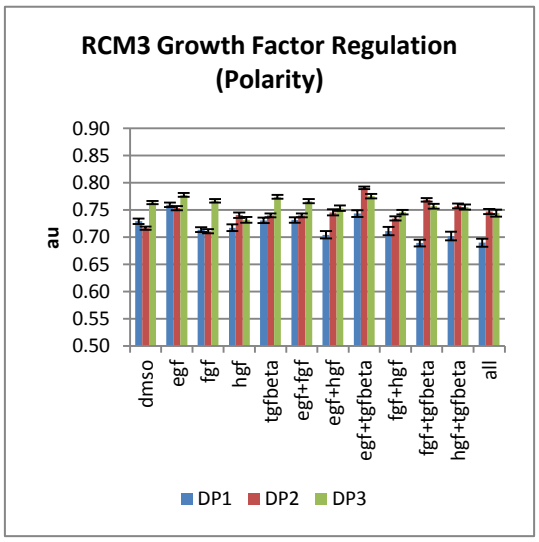
In this section, we introduce the persistence or motion linearity as another very important parameter for cell migration. Persistence is defined as a measure of how efficient the cells move [10]. The analysis result (cf. Figure 4-7) shows that HGF alone or in combination (except for HGF+FGF) does significantly enhance the migration speed of MTLn3 cells. The EGF only slightly affects the speed and interestingly FGF and TGF- β1 do not affect the speed at all. Furthermore, it seems that an increased velocity is always associated with an increased cell protrusion measured through extension [9]. So we can conclude that faster cells show an increased extension as a major feature. On the one hand, the EGF alone or combined with TGF- β1 always affects the motion linearity. On the other hand, the HGF alone does not change much of the motion linearity of the MTLn3 cells.



(a) velocity



(b) extension

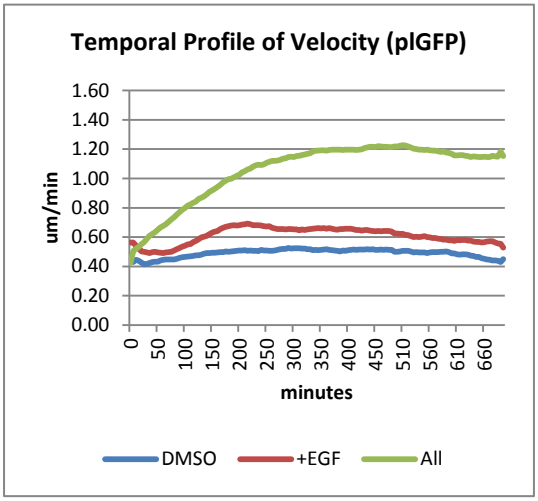


(c) polarity

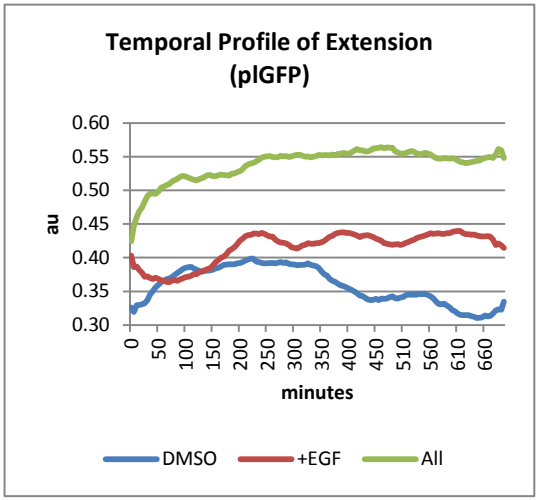
Figure 4-7 RCM3 Growth Factor Regulation plGFP

4.2.3. Temporal-Order Statistics

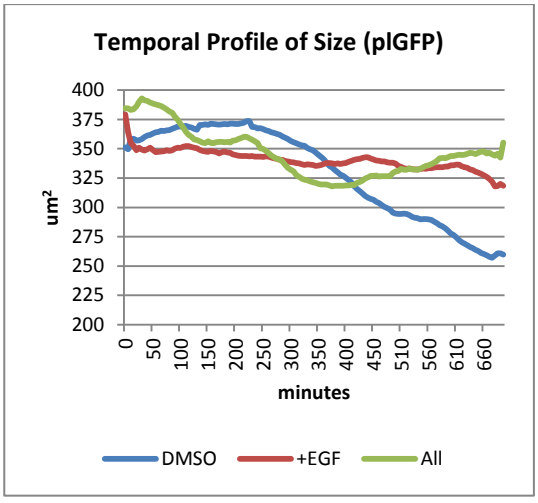
In many cell types, GFs have been reported to enhance the average migration speed. Especially in MTLn3 cells, EGF has been reported to increase the migration speed. The long term migration response of cells after exposure to EGF cannot be extrapolated from a known model. Basic first-order statistics do not reveal time related tendencies, temporal-order statistics [39] are further employed for the spatio-temporal analysis of cell migration. We plotted different parameters over the time-lapse (cf. Figure 4-8) and we ended up with descriptive time profile of the different stimulations. EGF stimluation did result in higher velocity through increased extension. Indeed, almost 1 hr after EGF stimulation the cells start to move faster until reaching a steady state (after approximately 5 hrs) of migration compared to the DMSO condition (negative control). On the other hand, a mixture of all GFs does greatly affect the migration speed of the MTLn3 cells since the temporal profile shows that the velocity increases continuously over the whole time-lapse sequence revealing a complex cascade of signaling that takes place at cellular level. This faster migration results in a continuous increase of cell extension.



(a) velocity



(b) extension



(c) size

Figure 4-8 temporal profile of plGFP

Cell tracks were generated and in our system it does reveal the heterogeneity in behavior within cell population. EGF stimulation indeed seems to increase the velocity only of a certain subpopulation, whereas the GF cocktail does affect all cells, resulting in a significant increase of the mean cell migration speed.

4.3. Conclusions and Discussion

Cell motility is an important event in many biological processes, such as tissue repair, metastatic potential, chemotaxis or analysis of drug performance. Cell migration and invasion are crucial aspects of tumor progression. It depends on the intrinsic capacity of the cells to move faster and more efficient. It is also depending on the tumor microenvironment that does produce all kind of cytokines or extra-cellular components. Understanding how tumor cells move and how their movement can be controlled by internal signaling pathways or external cues, in particular, contributes to our knowledge on tumor progression and metastasis formation. In order to obtain a better insight in the underlying processes leading to an efficient tumor cell migration, methods need to be developed that enable the study of migration at the individual cell level and in a high throughput fashion. Time-lapse imaging with fluorescent microscopy enables a thorough quantification of the cell dynamics at the level of migration. In this study, we successfully developed a computational approach that allows **(1)** reliable segmentation of migrating cells with a high degree of plasticity, **(2)** reliable tracking of fast and often dense migrating cells, **(3)** extraction of motility and morphology parameters over the time of imaging. In this chapter, we have illustrated a dedicated image analysis solution for HT/HC screen analysis combining robust **Watershed Masked Clustering (WMC)** segmentation algorithm and **Kernel Density Estimation (KDE) mean shift** tracking algorithm. This integrated solution produces an accurate profiling of cell migration behavior under the influence of growth factor regulation.

To ensure that our methodology was sufficiently competent for fast moving cells, we performed a pilot study on the migratory behavior of the rat breast carcinoma MTLn3 cells. They show a mesenchymal morphology and do not cluster with cells from the negative control. They are highly motile and propose a challenge for finding good segmentation and tracking solutions. In contrast to generic solutions, our dedicated solution of **WMC** segmentation and **KDE mean shift** tracking algorithm demonstrated a good performance in term of both algorithm efficiency and representation of the biology.

From the biological validation, we further conclude that robustness of both segmentation and tracking algorithm is a crucial factor behind successful analysis in a HT/HC screen study when cell responses from induced treatment are unknown. If the parameter set of an image analysis algorithm is tuned based on control cells demonstrating minimized cell protrusion and motility, with limited robustness, the same parameter set clearly is not optimized for treated cells such as +EGF cell that demonstrate an increasing cell protrusion and motility. The inflexibility not only produces errors in image analysis, but may lead to false conclusions.

On the workbench we resolve this dilemma via the employment of a pilot experiment consisting of (1) control condition in which no treatment is induced, (2) positive control condition with a treatment resulting in the maximum expected responses, and (3) negative

control condition with a treatment resulting in no responses. Parameters of image analysis algorithms and the robustness of selected solutions are verified using the pilot experiment before being applied to full-scale HT/HC screen study.

Chapter 5

A Study to the Dynamics of Matrix Adhesion

This chapter is based on the following publications

Yan, K., Le Dévédec, S., Van de Water, B., & Verbeek, F. J., “Automated Analysis of Matrix Adhesion Dynamics in Migrating Tumor Cells”, (in preparation)

S. Le Dévédec, B. Geverts, H. de Bont, K. Yan, F. J., Verbeek, A. Houtsmuller, and B. van de Water, “The residence time of focal adhesion kinase (FAK) and paxillin at focal adhesions in renal epithelial cells is determined by adhesion size, strength and life cycle status,” *Journal of cell science*, 2012.

Le Dévédec, S.E., Yan K., de Bont H., Ghotra V., Truong H.,Danen E., Verbeek F.J., and vande Water B.(2010), "A Systems Microscopy Approach to Understand Cancer Cell Migration and Metastasis", *Cellular and Molecular in Life Science*, Vol. 67, Issue 19, pp.3219-3240

Chapter Summary

In this chapter, we focus on assessing the performance of our image analysis solutions using a case study of the dynamics of a subcellular structure known as the matrix adhesion. We further divide the focus into two research questions:

1. Can we extract morphology and motility measurements from matrix adhesion dynamics and cell migration?
2. Can we identify correlation between measurements of matrix adhesion dynamics and cell migration?

Matrix adhesions are the closest contacts between the cell and the extra-cellular matrix through which both mechanical forces and regulatory signals are transmitted. To that end, cell migration can be seen as a cyclic process controlled by the assembly and disassembly of matrix adhesions. Therefore, a study of the correlation between MA dynamics and cell migration will provide an understanding of the molecular mechanism controlling cell migration behavior. For now, the study focuses on first extracting measurements describing both matrix adhesion dynamics and cell migration. These measurements are further used to reveal correlations between the morphology or the motility of MA dynamics and cell migration.

To achieve the current research questions, we will first extract measurements describing MA dynamics and cell migration using image analysis solutions discussed in Chapter 2 and Chapter 3. Furthermore, with these measurements, we apply dependency tests to identify potential correlations between MA dynamics and cell migration. Following the path of analysis, this chapter is divided into two major sections. In the section 5.1, the design of image acquisition, image analysis and data analysis will be addressed. In the section 5.2, the results in biological context will be discussed. The result of analysis shows that our solution confirms a number of known correlations observed in previous studies and reveals several yet unknown biological phenomena.

5.1. Workflow of Matrix Adhesion dynamics Analysis

Cell migration is an essential procedure involved in a number of processes and is especially important in cancer metastasis. In the cascade of cancer metastasis, an increasing in cell motility is crucial for cancer cells to invade the surrounding tissue. Matrix adhesions (MA) are the closest contacts between the cell and the extra-cellular matrix. Cancer cell migration can be seen as a cyclic process which is controlled by the assembly and disassembly of matrix adhesions. Therefore, the dynamics of matrix adhesion is very important [131] for the understanding of cell migration behavior, yet little is known about the molecular mechanisms that regulate adhesion dynamics and signaling during cell migration. In order to gain understanding of the relationship between cell migration and the dynamics of matrix adhesions (MA), we have developed an integrated approach consisting of image acquisition, image analysis and data analysis at both cellular and structural level:

1. For the image acquisition, epifluorescence and total internal reflection fluorescence (TIRF) microscopy are employed to respectively visualize components including the **cell body**, the **cell nucleus** and **matrix adhesions** (TIRF). This results in multi-channel time-lapse image sequences.
2. From these image sequences, measurements of cell migration and matrix adhesion dynamics are extracted using dedicated image analysis pipeline. These measurements represent a spatio-temporal quantification of the matrix adhesion dynamics in the migrating cell. Furthermore, a spatial model of the cell under migration is built, dividing the migrating cell into a number of characteristic regions that are subsequently used in the analysis. This cell body model is the key in our integrated approach to find correlation between matrix adhesion dynamics and cell deformation.
3. From the measurements, dependency tests are applied to find significant correlations between matrix adhesion dynamics and cell deformation.

Each step will be further illustrated in: experiment preparation and image acquisition (cf. § 5.1.1), (2) image analysis (cf. § 5.1.2), and (3) data analysis (cf. § 5.1.3).

5.1.1. Experiment Preparation and Image Acquisition

Material Preparation

The H1299 cell model, which ectopically expresses the GFP-paxillin matrix adhesion marker, is the well described lung carcinoma cell-line extensively used in cell migration assays [132]. H1299 cells (ATCC-CRL-5803) were cultured in RPMI (GIBCO, Life Technologies, Carlsbad, CA, USA) supplemented with 10% FBS (PAA, Pasching, Austria) and 100 International Units/ml penicillin and 100 µg/ml streptomycin (Invitrogen, Carlsbad, CA, USA). CELLview glass bottom dishes with four compartments were coated with 10 µg/µl fibronectin (Sigma Aldrich) for 1 hr at 37^{o}C. H1299/GFP-paxillin cells were seeded on glass bottom dishes and grown at 37^{o}C overnight. For random cell migration assays, phenol-red (pH indicator) free culture medium was used. Cells were maintained in a 5% CO_2 humidified chamber at 37^{o}C. Following this protocol, the cells are believed to be incubated with a minimum level of interference.

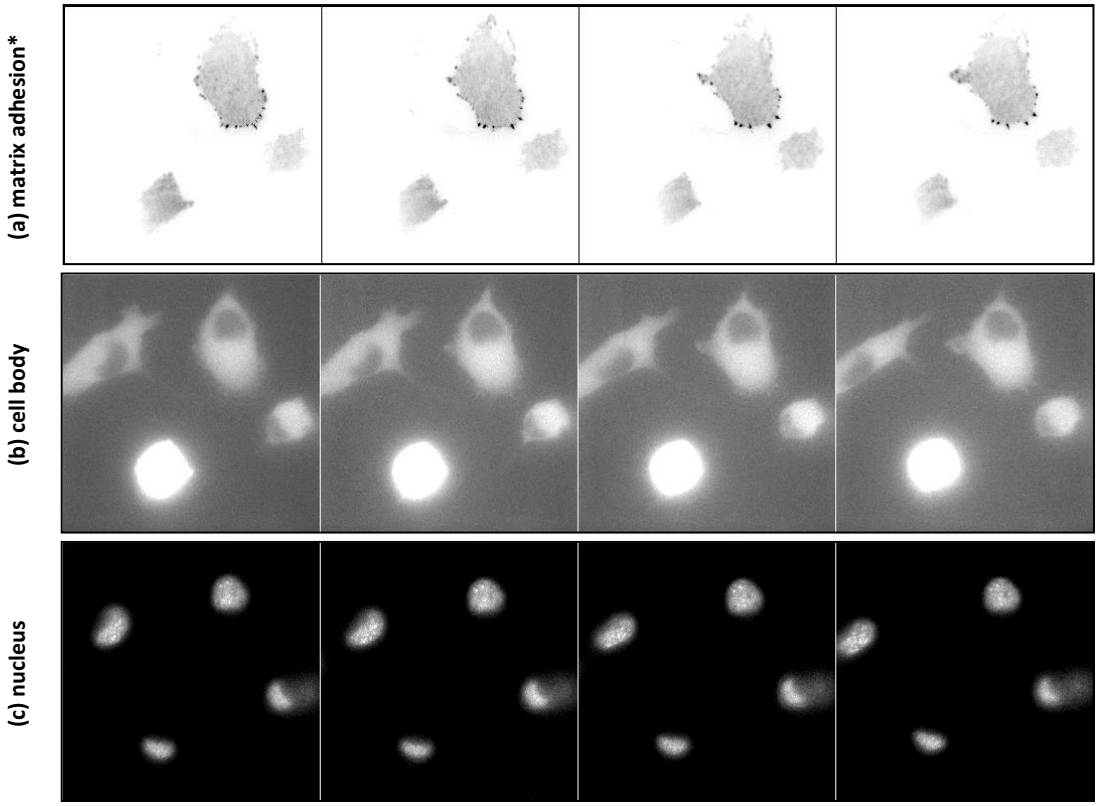


Figure 5-1 multi-channel imaging for MA dynamics analysis, image montage at time [0 min, 16 min, 24 min, 32 min]; *signal reversed for visibility

Image Acquisition

In order to quantify both cell migration and MA dynamics, we captured time-lapse image sequences for each component using different imaging techniques. Prior to imaging, nuclei were labelled with 100ng/ml Hoechst 33342 [133] in culture medium for 45 min. H1299/GFP-paxillin cells were imaged using a Nikon TiE2000 microscope equipped with a Perfect Focus System with 5% CO_2 delivery to the sample dish. Images were acquired with a 60x oil objective (1.49 NA, 0.12 WD) and the image acquisition was controlled by NIS Elements (Nikon). The imaging setting is defined as follows:

1. Matrix adhesions were captured using TIRF imaging with a 488nm laser line over a period of two hours, in one minute intervals. The TIRF imaging was used to detect the signal of GFP-paxillin that is localized in the first 80 nm of the cells where the MAs are found (cf. Figure 5-1a). The TIRF imaging is in particular useful for visualizing objects with a fast turnover and small size [134]. Additionally, it provides a good signal-to-noise ratio for capturing dynamic matrix adhesions [135].
2. The total signal of the cytoplasmic GFP, which represents the whole cell body information, is detected using wide field microscopy (cf. Figure 5-1b). The cytoplasmic GFP signal was acquired every three minutes over a period of two hours using wide-field fluorescence. (3)
3. At the same time interval (3-mins) and duration of 2-hrs, the Hoechst [133] signal for detection of the nucleus (cf. Figure 5-1c) was acquired using wide-field fluorescence.

Image Format

The raw images are stored in the ND2 format; a commercial self-contained image data storage format proposed by NIKON. Due to its patented structure, ND2 format is not an open-source storage format and can only be accessed via the Windows-based native API provided by NIKON, consequently the ND2 format is not convenient for cross-platform data transportation. Therefore, the ND2 format is converted into TIFF image format. The conversion is conducted as follows: for each time point in a time-lapse image sequence, an image from the separated channels is stored as a 16-bit TIFF file (cf. Figure 5-2). As a result, the ND2 file is converted into a collection (time-lapse image sequences) of individual 16-bit TIFF files.

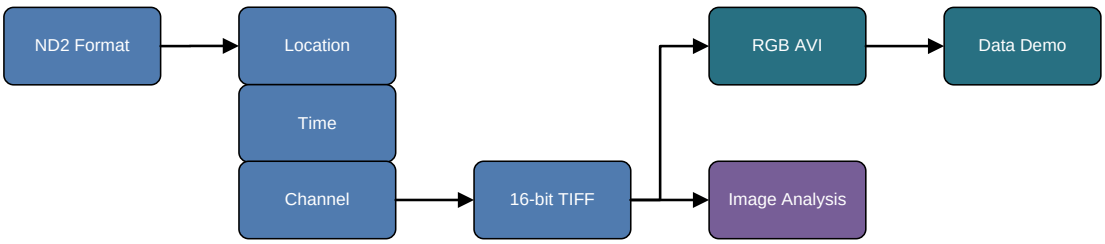


Figure 5-2 2D+T ND2 to TIFF conversion

The converted TIFF files are organized and stored in a structure illustrated in Figure 5-2. The total length of image sequence is 120 ~ 240 images depending on the size of observation window. For data visualization, a copy of the 16-bit TIFF image is combined into multi-channel RGB AVI format by rescaling the 16-bit TIFF image into 8-bit. The RGB AVI format conversion is accomplished with ImagePro macro [136].

In conclusion, by using two different fluorophores (GFP and Hoechst) and two different imaging techniques (TIRF and wide-field fluorescence), we are able to capture three different sets of information: the cell nucleus (wide-field/Hoechst), cell body (wide-field/GFP) and matrix adhesions (TIRF/GFP) respectively (cf. Figure 5-1).

Different from the cell migration study in Chapter 4, the current imaging setting is not suitable for a high-throughput setting due to its time resolution. Image analysis is however still required to extract measurements of MA dynamics and cell migration. In the next section, we will discuss the image analysis pipeline.

5.1.2. Image Analysis

The image analysis pipeline answers to the first research question, namely to extract measurements of MA dynamics and cell migration. It is the essential step to convert images into measurements by using image segmentation and object tracking. In correspondence with the imaging setting of MA dynamics, the image analysis solution consists of three parts (cf. Figure 5-3). The selection of image segmentation and object tracking solution is based on the characteristic qualities of each of the channels.

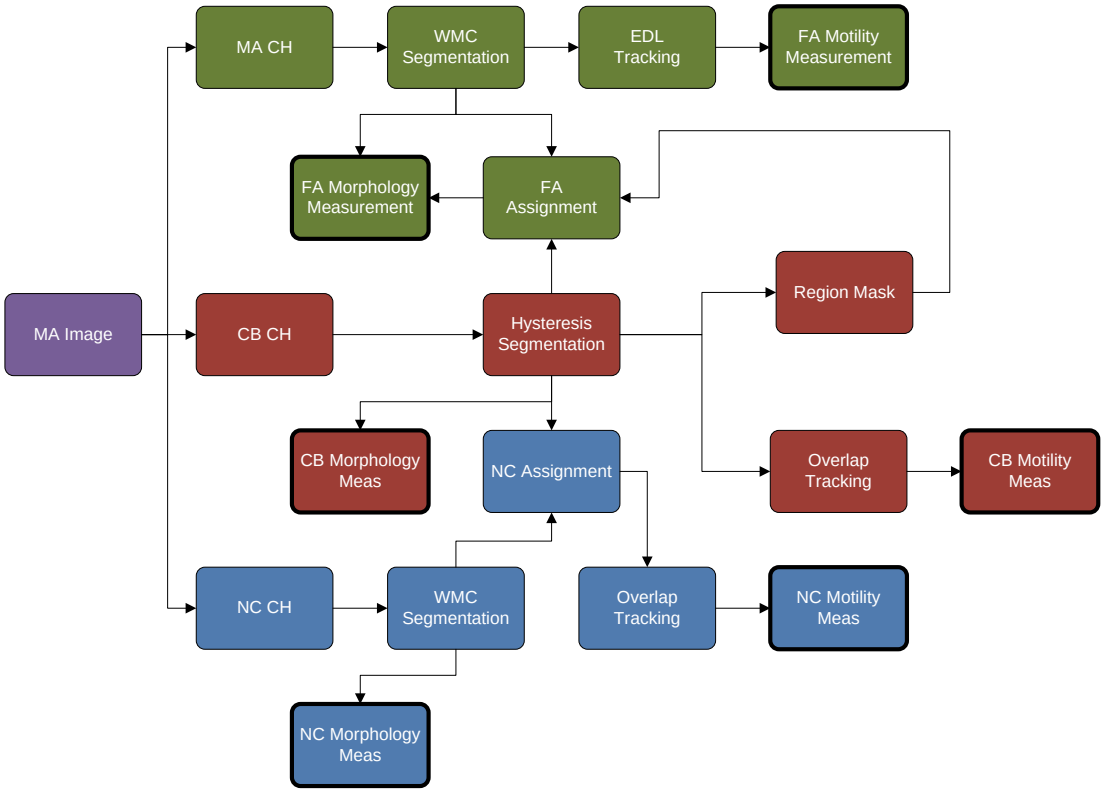
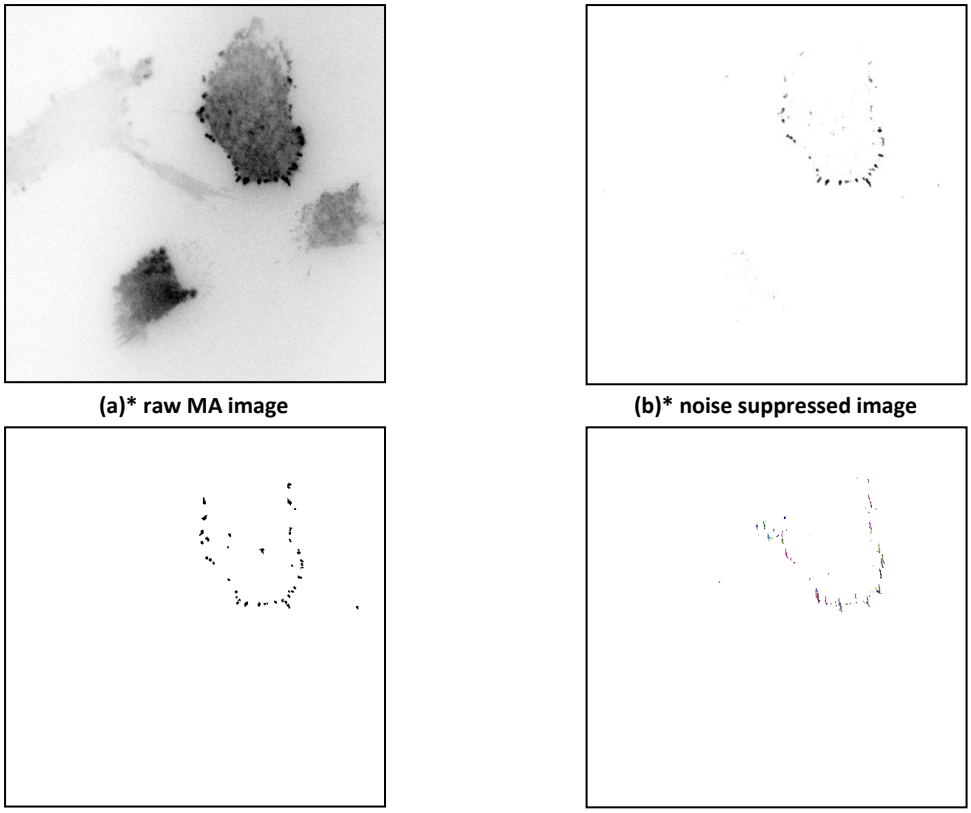


Figure 5-3 workflow of MA image analysis

5.1.2.1. Image Analysis for Matrix Adhesion (MA) Channel

The image analysis of the matrix adhesion channel is illustrated in Figure 5-3 [green]. With TRIF microscopy, the GFP signal from the cytoplasm and the MAs are both captured. As a result, the MA channel (cf. Figure 5-4a) contain two layers of signals including the brighter MA signal and comparatively darker cytoplasmic signal. In order to extract only the MA, the cytoplasmic signals must be removed from the image. By treating the cytoplasmic signal as a part of the background, we extract the major MA signal using a combination of Gaussian blurring filter and rolling ball background subtraction algorithm [137]. Empirically we established both the σ of Gaussian filter and the kernel size r_r of the rolling ball filter. The choice of $\sigma = 1$ is just larger than the average radius of MAs. The choice of $r_r = 2$ is twice the average radius of MAs. By applying both filters to the MA channel, we can suppress the cytoplasmic signal [35] while preserving the brighter MA signal (cf. Figure 5-4b) below the radius of the rolling ball. The resulting image only contains the dominant MA signal; the WMC segmentation algorithm will be used to create binary masks for the MAs (cf. Figure 5-4c).

The motivation for the WMC segmentation algorithm is similar to the motivation given in Ch. 4. The MA image (cf. Figure 5-4b) contains a large intensity variation possibly due to the Z-position of each MA. Moreover, the assembly and disassembly procedure of MAs will lead into a temporal change in intensity values of the same MA. To that end, the WMC segmentation algorithm is a good choice since it is designed to adapt the threshold based on local intensity information. Each object in the binary mask is labeled and tracked using the EDL tracking algorithm (cf. Figure 5-4d). As described in Chapter 3, the EDL tracking algorithm is designed for this particular kind of study.



(a)* raw MA image (b)* noise suppressed image

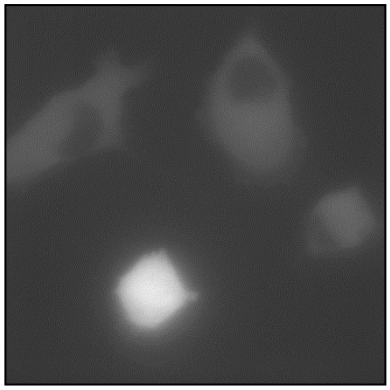
(c)* binary mask (d)* tracked trajectories

Figure 5-4 raw MA image and intermediate results of MA channel analysis (*) signal reversed for visibility

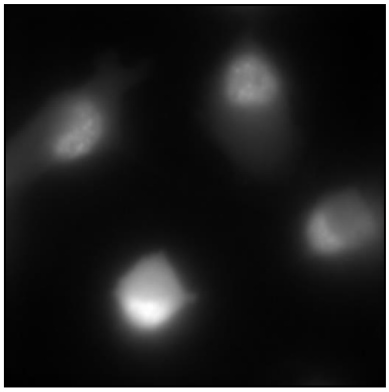
5.1.2.2. Image Analysis for Cell Body Channel

The workflow for the analysis of the cell body (CB) channel is illustrated in Figure 5-3. At the given magnification level, the cell body channel alone contains complex textures and multiple maxima that are sensitive to overcutting when segmented with WMC. To overcome potential overcutting, the cell body channel (cf. Figure 5-5a) is combined with the NC channel (cf. Figure 5-5b) to introduce a more precise definition of the maxima for object separation. Since each cell body can only have one nucleus, the NC channel is a perfect seed channel for segmenting the cell body channel using WMC. The combined image is applied with a Gaussian blurring filter ($\sigma = 3$) which is just larger than the average diameter of the MA, thereby suppressing the potential local maxima from the MA signal. Finally, the blurred image is segmented using WMC algorithm (cf. Figure 5-5c). Subsequently, each object is labeled and tracked using the overlap tracking algorithm (cf. Figure 5-5d). Objects touching the image border are discarded since they only give partial information.

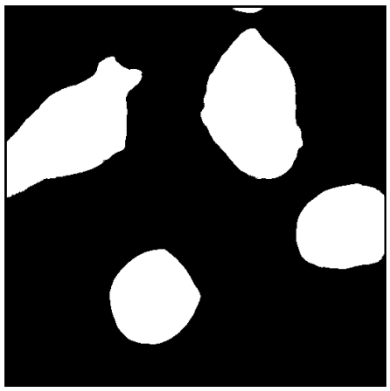
In order to analyze the MA dynamics with respect to cell behavior, we have modeled the cell in regions. From the binary mask of the cell body channel, six functional regions are defined within each mask. These regions are hierarchically related and the mixture of these relationships is illustrated in Figure 5-6. The definitions for each of the functional region is described in the literature [138][112]. The functional regions are derived from major episodes in the cell body deformation during cell migration.



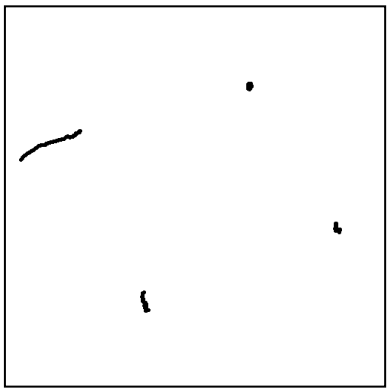
(a) raw cell body image



(b) cell body + nculeus image



(c) binary mask



(d)* tracked trajectories

Figure 5-5 raw cell body image and intermediate results of cell body channel analysis (*) signal reversed for visibility

The MAs are assigned to each functional region so that, in each episode of cell deformation during migration, the study of MA dynamics can be assessed. The control mechanism behind MA-cell body correlation is often interconnected; therefore the study of difference between MA-cell body correlation models for the different functional regions may reveal new insights. Next, we will explain the regions:

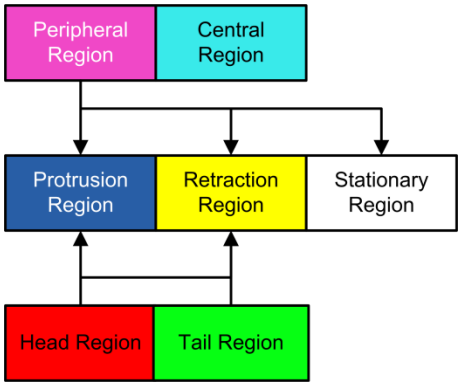
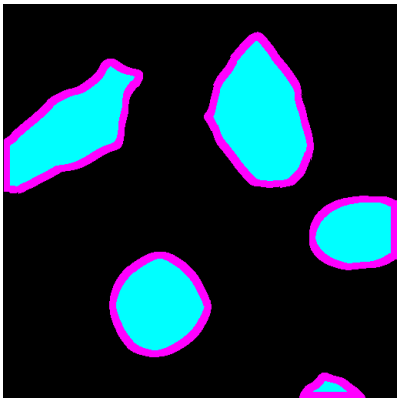
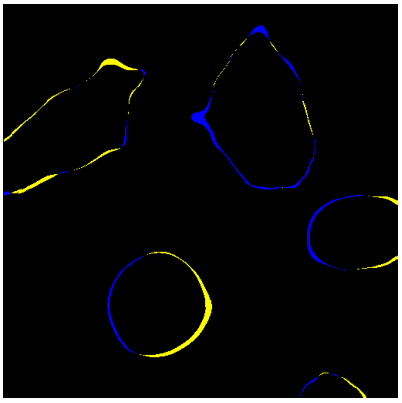


Figure 5-6 relation between each region

1. **Peripheral Region (P)**: the cell membrane region at the border
2. **Central Region (C)**: the inner cytoplasm region around nucleus
3. **Protrusion Region (PR)**: lamellipodium protrusion formation
4. **Retraction Region (RE)**: cell body retraction
5. **Front Region (F)**: the whole leading edge of cell body during migration
6. **Back Region (B)**: the rear edge of cell body during migration



(a) central [cyan] & peripheral [magenta]



(b) protrusion [blue], retraction [yellow]

Figure 5-7 functional regions derived from cell body masks

Peripheral and Central

The **peripheral region (P)** and **central region (C)** (cf. Figure 5-7a) are two geometrical compartments in the cell body defined as follows:

1. The **peripheral region** is the outer ring (near membrane) of the cell body mask
2. The **central region** is the remaining area (inner part) of the cell body mask.

Hereby, the **peripheral region** and **central region** are correspondingly defined as the outer ring and intra region of the cell body (cf. Equation 5-1) based on a user-defined width of that ring. In Figure 5-7a, a peripheral region [magenta] of a width of 20 pixels is illustrated; the width corresponds with image resolution and is derived from empirical observation. The peripheral region (P) and the central region (C) are mutually exclusive, this can be written as:

$$(P \cup C = cell\ body) \wedge \left(P = \overline{C}\right) \qquad \textbf{Equation 5-1}$$

The definition of the peripheral and central region allows the study of MA dynamics near the cell border. From the literature [138][139], it has been reported that a rapid MA formation at the peripheral region is strongly associated with both cell motility speed and signaling.

Protrusion and Retraction

The **protrusion region (PR)** and the **retraction region (RE)** (cf. Figure 5-7b) are two regions derived from both geometrical and temporal information of the cell body. They are defined as the shape variation between the cell body masks from two consecutive time points as follows:

$$PR = b_{i+1} \cap \overline{b_i} \qquad \textbf{Equation 5-2}$$

$$RE = b_i \cap \overline{b_{i+1}} \qquad \textbf{Equation 5-3}$$

, where b_i and b_{i+1} is the peripheral region of a cell body in i^{th} and $(i+1)^{th}$ frame. It is reported [138][15] that MA dynamics in the protrusion and retraction region, in terms of lifetime and turnover, is strongly associated with cell motility and migration polarity.

Front and Back

The **front region (F)** and **back region (B)** (cf. Figure 5-8c) are high-level perceptualizations of cell body regions. They represent the leading area and rear area of a cell migration. From recent studies [138][15][140], one can deduce the opinion that cell migration can be described by a combination of adhesion formation and cytoskeleton formation at both leading area and rear area. Empirical observations [138][140] suggest that lamellipodia protrusions are first

formed by expanding the cell body structure at the leading area. Cell body adhesions, i.e. matrix adhesions, are assembled in the protrusions to push the extracellular matrix to attach to the substrate surface. Meanwhile, at the rear area adhesions are gradually disassembled and release extracellular matrix from substrate surface. Some theories [140] also pointed out that the disassembly of cell body adhesions also provides pushing forces during migration.

To study how the matrix adhesion dynamics in these two areas are connected to cell migration, we define the **front region** as the cell body region aligned with the **leading edge direction** while the back region as the cell body region is aligned with the **rear edge direction**. The leading edge direction (cf. Figure 5-7b) is defined as the direction of the joint force, assuming each protrusion region as pulling force [138][140] and each retraction region as pushing force [138]. Similarly, the rear edge direction (cf. Figure 5-7b) is defined as the opposite direction of the joint force, assuming each retraction region as pushing force. Given that:

1. The pixels $P_{p(i)}(t)$ belong to the protrusion region
2. The pixels $P_{r(j)}(t)$ belong to the retraction region
3. The $P_x(t)$ and $P_y(t)$ represents the *x*-*y* coordinate of the pixel $P(t)$ at time point *t*
4. The $NC_x(t)$ and $NC_y(t)$ denote the center of mass of nucleus at time point *t*
5. The α is the direction of nucleus positional shift

Then, the definition of the leading edge direction γ_F and the rear edge direction γ_R can be derived as follows:

$$\gamma_F = atan\left(\frac{\sum_{t=1}^{n}\omega(t)\cdot\left[NC_y(t)-P_{p(i)y}(t)\right]}{\sum_{t=1}^{n}\omega(t)\cdot\left[NC_x(t)-P_{p(i)x}(t)\right]}\right) \quad \textbf{Equation 5-4}$$

$$\text{, where } \omega(t) = \left|\frac{\beta\left(P_{p(i)}(t)\right)-\alpha}{\pi}\right| \quad \textbf{Equation 5-5}$$

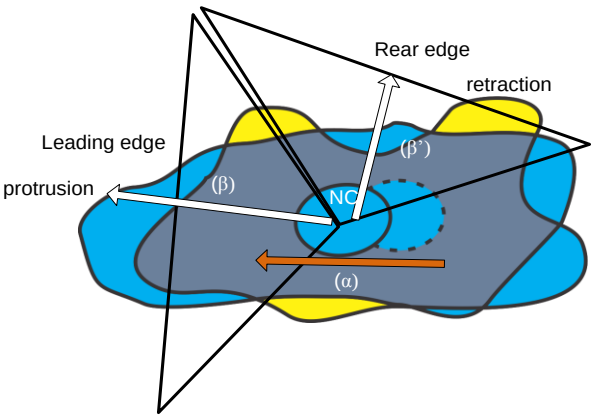
$$\text{, where } \beta\left(P_{p(i)}(t)\right) = atan\left(\frac{NC_y(t)-P_{p(i)y}(t)}{NC_x(t)-P_{p(i)x}(t)}\right) \quad \textbf{Equation 5-6}$$

$$\gamma_R = atan\left(\frac{\sum_{t=1}^{n}\omega'(t)\cdot\left[NC_y(t)-P_{r(j)y}(t)\right]}{\sum_{t=1}^{n}\omega'(t)\cdot\left[NC_x(t)-P_{r(j)x}(t)\right]}\right) \quad \textbf{Equation 5-7}$$

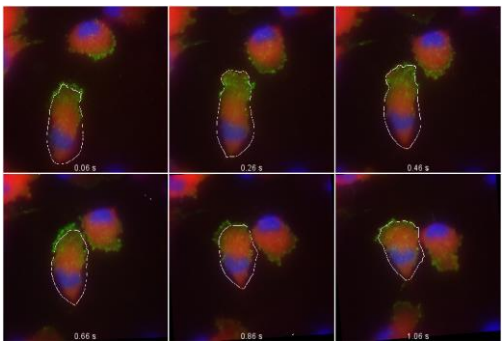
$$\text{, where } \omega'(t) = \left|\frac{\beta'\left(P_{r(j)}(t)\right)-\alpha}{\pi}\right| \quad \textbf{Equation 5-8}$$

$$\text{, where } \beta'\left(P_{r(j)}(t)\right) = atan\left(\frac{NC_y(t)-P_{r(j)y}(t)}{NC_x(t)-P_{r(j)x}(t)}\right) \quad \textbf{Equation 5-9}$$

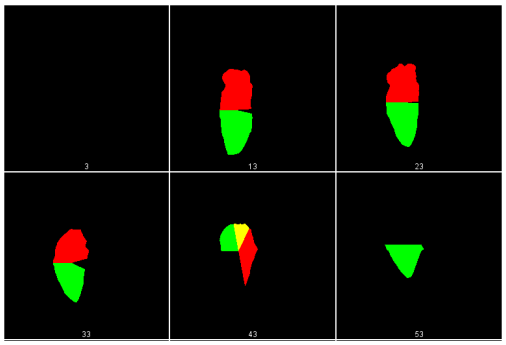
The weight factors $\omega(t)$ and $\omega'(t)$ are introduced under the condition that the direction β and β' will contribute more to the pulling/pushing force, if the β and the β' are aligned with the direction of nucleus position shift. By this notion, we intend to damp the pulling/pushing forces caused by small protrusions or retractions; as these are believed not to contribute to the migration [138].



(a) principle of front and back region recognition



(b) the selected cell



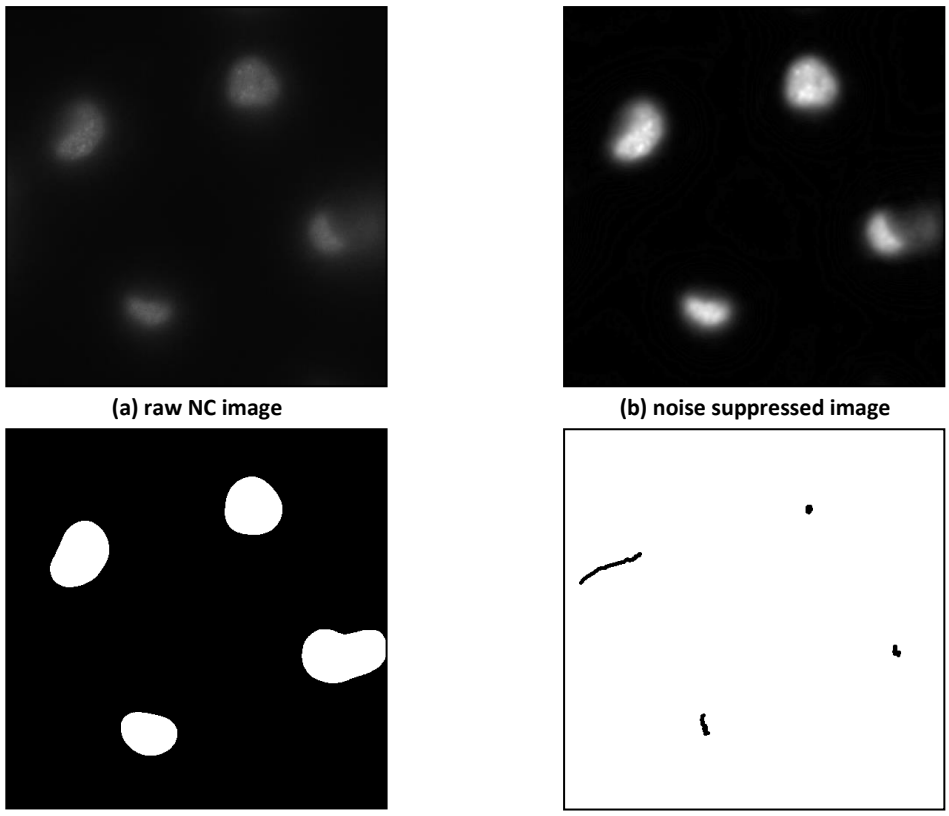
(c) front region [red] and back region [green]

Figure 5-8 front and back region recognition

Each MA is located in either of the functional regions; in this manner the MA obtains a region label from the region model. A MA can be labeled with multiple regions, for example, a MA can be in protrusion region and at the same time in head region. Some of the regions are, however, exclusive. For example, a MA cannot be simultaneously assigned to the peripheral and the central region since the definition of peripheral and central region is mutually exclusive. Using definitions of the functional regions as grouping criteria, the per-region analysis of MAs allows MA dynamics to be linked directly to local cell deformation or migration.

5.1.2.3. Image Analysis for Nucleus Channel

The analysis of the nucleus (NC) channel (cf. Figure 5-9a) is illustrated in Figure 5-3. A Gaussian blurring filter is first applied to the image to smooth the intensities and remove noise [28]. Here we choose a $\sigma = 4$ for the Gaussian filter that is sufficient to suppress the Poisson noise (cf. Figure 5-9a) [28] and create a more smooth intensity landscape within the nucleus (cf. Figure 5-9b). The blurred image (cf. Figure 5-9b) is segmented with WMC segmentation algorithm. The binary mask (cf. Figure 5-9c) is labeled and subsequently tracked using the overlap tracking algorithm (cf. Figure 5-9d).



(a) raw NC image (b) noise suppressed image

(c) binary mask (d)* tracked trajectories

Figure 5-9 raw NC image and intermediate results of NC channel analysis (*) signal reversed for visibility

5.1.2.4. Phenotypical Quantification

With image analysis (cf. Figure 5-3) measurements describing both MA dynamics and cell migration are extracted. Apart from the morphology and motility measurements [9], several correlation measurements are also introduced to describe morphological or motile association between cell body deformation and MA dynamics (cf. Figure 5-10).

In order to extract the correlation measurements, objects from different channels are first related according to parent-child relationships. MAs are assigned to cell bodies as children based on the minimum distance between contour of cell body and MA. Nuclei are assigned in a child relation to the cell bodies based on the overlapping ratio between NC and cell body. From these two parent-child relationships, the following measurements are defined (cf. Table 5-1):

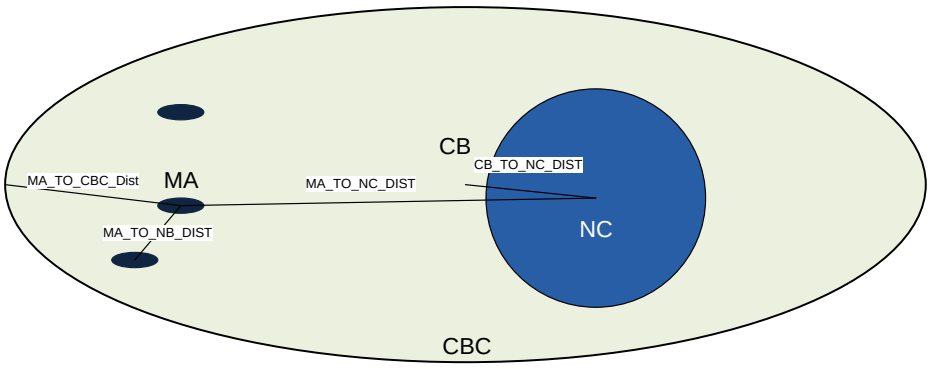


Figure 5-10 visualization of correlation measurements described in the Table 5-1 (MA: Matrix Adhesion; CB: Cell Body; NC: Nucleus; CBC: Cell Body Contour)

Table 5-1 definition of correlation measurements in current study

Correlation measurements	
Table Name	**Description**
MA_TO_NC_DIST	The Euclidian distance between the center of mass of a MA to its NC, $MA \ \& \ NC \ \in CB$
MA_TO_CB_DIST	The minimum Euclidian distance from a MA to the nearest pixel of cell body contour according to MA label
CB_TO_CB_DIST	The minimum shifting distance between two consecutive cell body contours
CB_TO_NC_DIST	The shortest Euclidian distance between the center of mass of one NC to the cell body which it belongs
MA_TO_NB_DIST	The Euclidian distance between a MA to its nearest MA in the same cell body

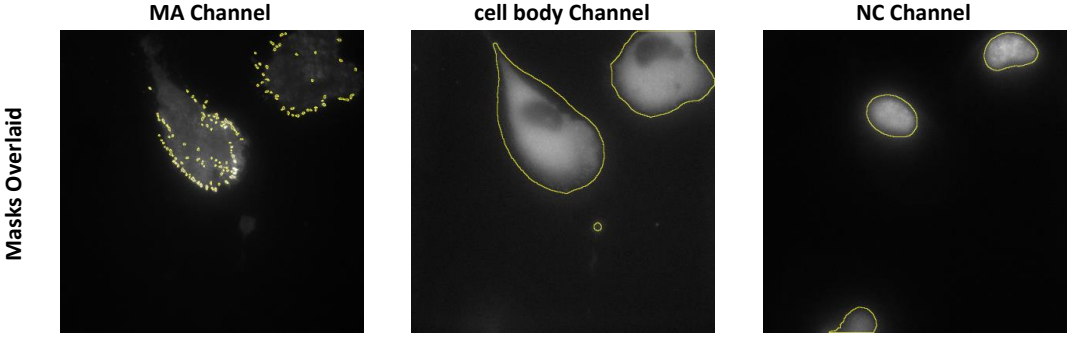
5.1.3. Data Analysis

With measurements extracted from image analysis, we address the second research question, namely to identify correlated measurements between MA dynamics and cell body. To reveal correlation knowledge from in the data, an unsupervised correlation analysis is further verified based on expert observations and literature [36][138][141][112][111]. In the current implementation, the correlation analysis employs a statistical approach namely the Pearson cross-correlation [142] for potential linear correlation. The Pearson product-moment correlation, or simply Pearson correlation, is widely used as a measurement of the strength of linear dependence between two variables [124][143]. Here we calculate the correlation between measurements of MA and cell body in a pairwise fashion. By extracting significantly correlated measurements between cell bodies and MAs in each functional region, we hope explain numerical causality between MA dynamics and cell migration. Moreover, we hope to identify different patterns on how the MA dynamics in different functional regions is correlated to cell behaviors.

User Verification

The user verification step of MA dynamics is performed in a similar fashion as the cell migration study (cf. Ch. 4). With overlay information (cf. Figure 5-11), researchers are asked to assess the following aspects of the general performance of image analysis for each channel:

1. Image segmentation
2. Object separation
3. Object tracking



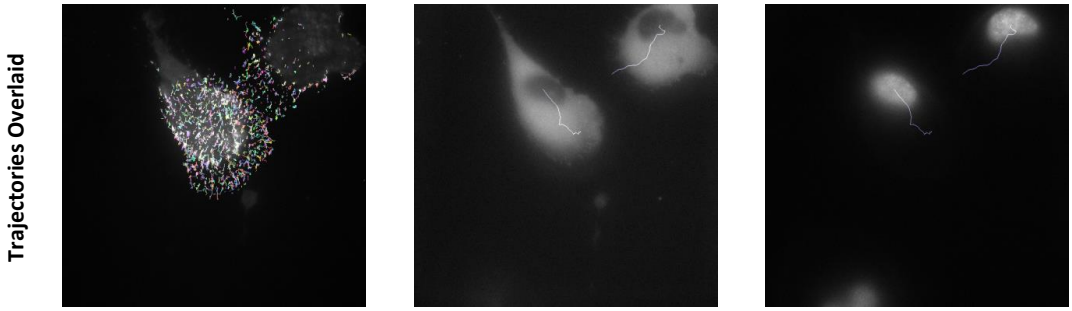


Figure 5-11 image with overlaid information for manual verification

5.2. Phenotypical Correlation Study of the Live Cell

Cell migration is a well-orchestrated event that consists of several phenotypical stages believed to be controlled by the assembly and disassembly of matrix adhesions. The study of how MAs controls cell phenotypes is important for the understanding of molecular control mechanisms behind cell migration. In this case study, we attempt to comprehend such correlation by introducing a high-content analysis of a cell model consisting of different functional regions. In section 5.1, we have demonstrated the possibility to quantitatively identify and verify potential correlations between the dynamics of MAs and cell migration. In this section, we will illustrate several correlations that are revealed and verified by our analysis.

Matrix Adhesion Lifetime and Cell Migration Velocity

A total collection of 29 time-lapse image sequences are captured using image acquisition settings described in §.5.1.1. All cells touching the image border or that are only partially present are discarded since they do not provide complete information on cell behavior and cannot be used to extract correct measurements of cell velocity or shape deformation. In this manner, there are 43 valid cells remaining for further analysis. Using the unsupervised K-means clustering algorithm, these 43 cells are divided into a low-motile (cf. Figure 5-12) and a high-motile class (cf. Figure 5-13) based on their migration velocity. With the two motility groups, we intend to extract major differences between MA dynamics. These major differences in MA dynamics are potentially candidates for the correlation modeling procedure since they are most likely to be associated with control mechanism of cell migration.

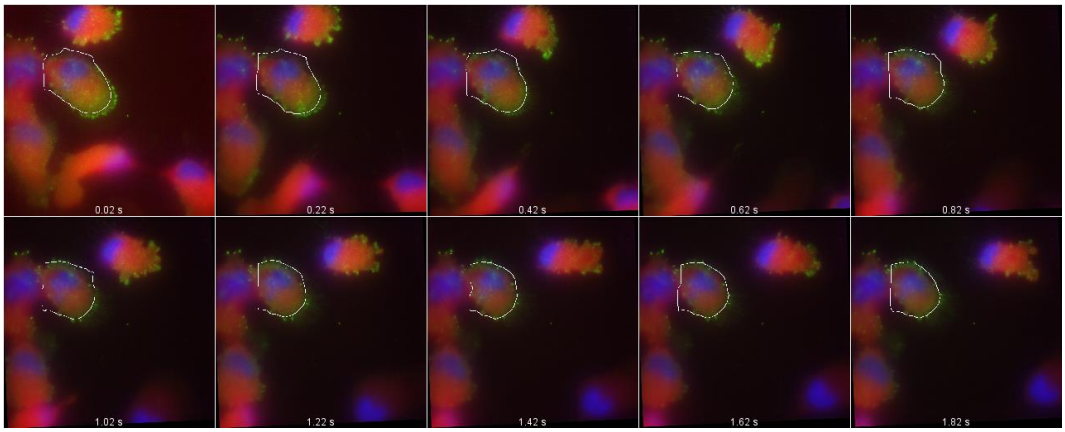


Figure 5-12 montage of low-motile cell

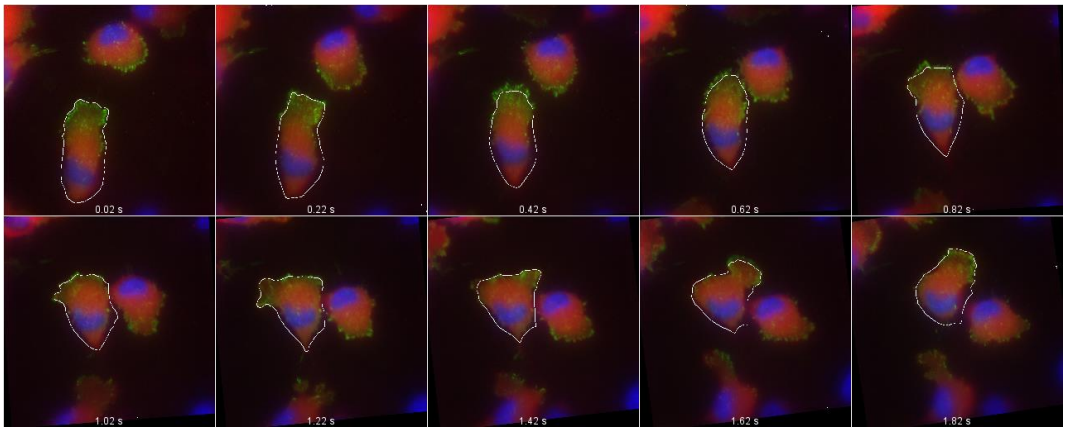


Figure 5-13 montage of high-motile cell

For cells in each velocity group, their MAs, in total 896, are further divided into eight subpopulations based on their labels from cell body functional regions:

1. MA Protrusion (MA PR): MAs located in the protrusion region of the cell body
2. MA Retraction (MA RE): MAs located in the retracting region of the cell body
3. MA Peripheral (MA P): MAs located in the peripheral region of the cell body
4. MA Central (MA C): MAs located in the central region of the cell body
5. MA Front: MAs located at the front of the migrating cell
6. MA Back: MAs located in the back of the migrating cell
7. MA Front PR: MAs located in protruding regions at the front of the migrating cell
8. MA Back RE: MAs located in retracting regions in the back of the migrating cell

MAs in each local cell region are separately analyzed. The Kolmogorov–Smirnov (K-S) test [124][144] is used for comparing measurements from each region since none of the MA or cell body measurements fits a normal distribution (based on Lilliefors normality test). The result of the K-S test with 95% confidence interval (cf. Figure 5-14) shows that:

1. In general, MAs in high-motile cells display a shorter lifetime compared to MAs in low-motile cells, suggesting that MA lifetime is correlated to cell velocity.
2. The difference between MA lifetime in peripheral and central region is larger in low-motile cells compared to the high-motile cell, suggesting that a shorter lifetime of peripheral MAs is necessary for a higher cell motile.
3. Surprisingly, the lifetime of central MAs is the lowest and does not differ between low-motile and high-motile cells.
4. In high-motile cells, the MA lifetime is always longer in the retracting region than in the protruding region.
5. The difference in MA lifetime between protrusion and retraction regions is absent in low-motile cells which may be an explanation for lower migration polarity.

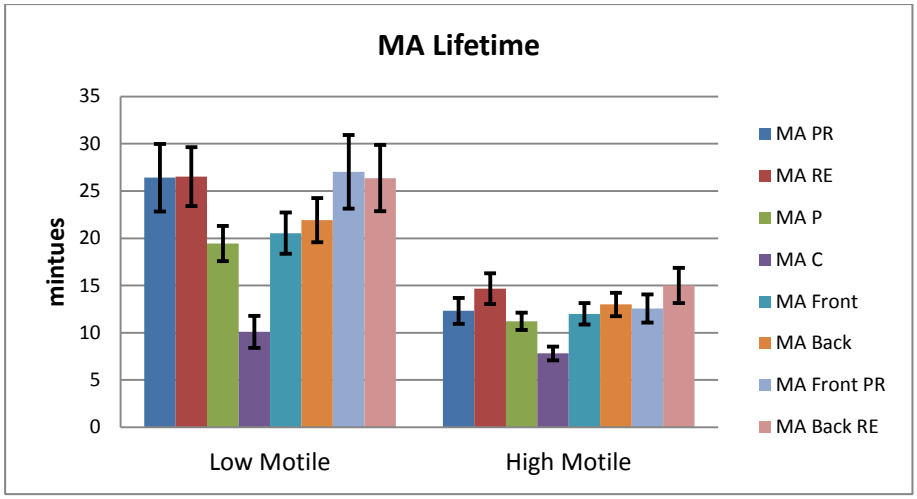


Figure 5-14 MA lifetime variability given the difference in cell velocities

Unsupervised Correlation Discovery

In order to investigate all potential correlations, we implemented an automated solution (cf. §.5.1.3) using Pearson correlation analysis [124][143]. The Pearson correlation analysis is a popular measurement of linear dependency between random data. Moreover, in order to test whether correlation is statistically significant, the Pearson correlation analysis transforms the problem into a one-sample test with a bivariate distribution. In other words, it is tested whether the sampled data belongs to a hypothetical population with the same mean and standard deviation of the sampled data. Heatmap visualizations of the *p*-values of the correlation test between measurements of MAs in the Front-PR and cells are depicted in Figure 5-15 and Figure 5-16. Each row represents a phenotypical measurement of cell migration and each column represents a phenotypical measurement of MA dynamics. The heatmap visualization provides a fast overview of all potential correlations.

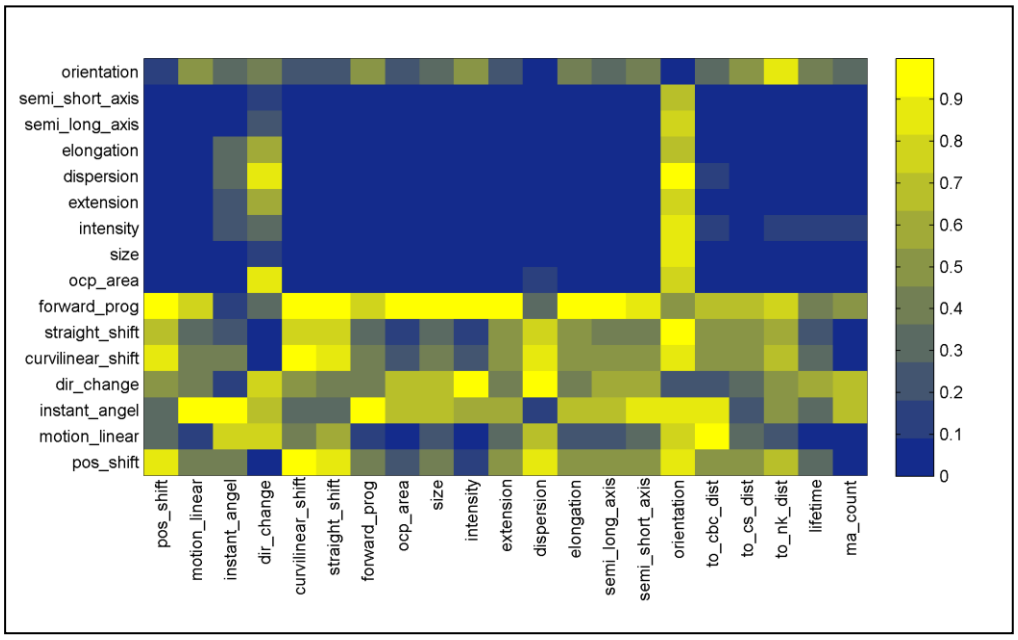


Figure 5-15 the heatmap of the p-value of correlation test between all cell (column) and MA Front PR (row) phenotypical measurements

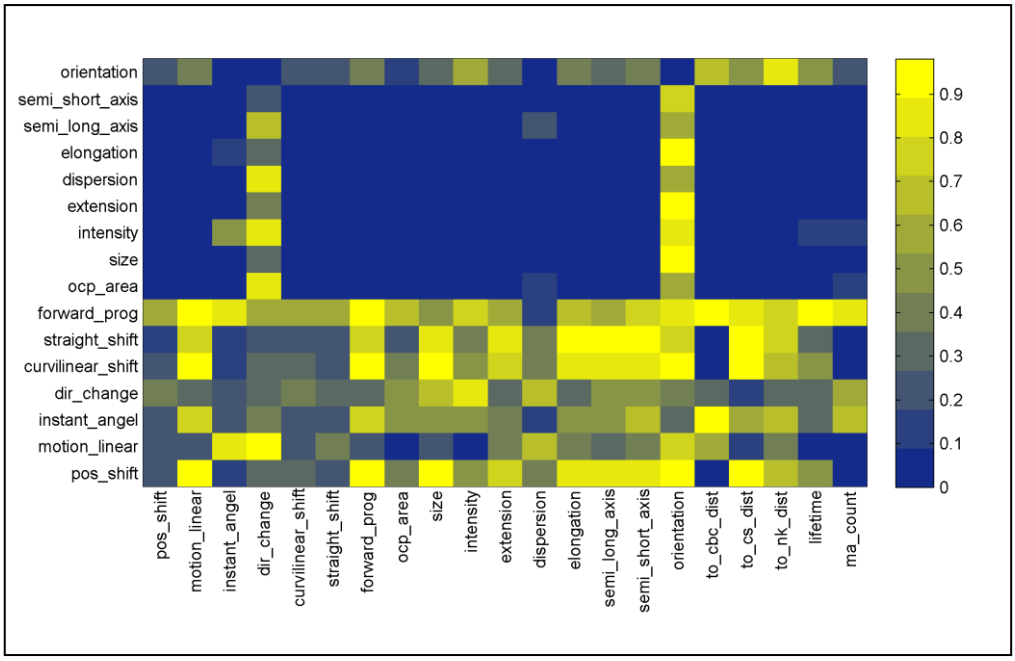


Figure 5-16 the heatmap of the p-value of correlation test between all cell (column) and MA Back RE (row) phenotypical measurements

By exploring the heatmaps of Figure 5-15 and Figure 5-16, it is clear that cell shape changes such as elongation or protrusion formation are highly associated with nearly all MA phenotypical changes in these two regions. However, cell motility patterns such as velocity or polarity are have a strong correlation with only a few MA measurements such as MA count in Front PR or MA to cell body distance in Back RE.

5.3. Conclusion and Discussion

In this case study, we have developed a set of image and data analysis solutions for the quantification of matrix adhesion dynamics and cell migration. First, the image analysis converts high-content image data captured (cf. § 5.1.1) into characteristic measurements for matrix adhesion dynamics and cell migration. From the measurements, we further introduce an automated data analysis solution to reveal correlations between the morphology and the motility of matrix adhesion dynamics and cell phenotypes. At the current stage, the combined solution can well provide the possibility to increase further understanding on the regulation of MAs and how this affects cell migration. Moreover, it paves the way to a numerical expression of the control mechanism behind cell migration.

There are several interesting issues that can be addressed in future studies:
From image analysis perspective: (1) when cells are clustering with each other, the border region is becoming less identifiable; thus results in a more complex cell separation. (2) The MA has a very small size (5~10um) and short life time (3~15 frames), which does not always provide sufficient information to train a motion model for object tracking. (3) Longer exposure to laser may cause cell apoptosis while a reduction of temporal-resolution image most certainly results in loss information on morphology changes. Thus, it leads into bias in measurements that cannot be easily detected in data analysis.

From data perspective: (1) Pairwise linear or monotonic correlation can be easily detected. However, it has not yet been elaborated on how to extract more complex correlation model from the measurements. (2) Since current measurements are derived from empirical observations, some biological phenomena such as the length of a protrusion may be overlooked. Yet, it is unclear how to define new measurements that capture these biological phenomena.

Chapter 6

Reasoning over Data in HT/HC Management

This chapter is based on the following publications

Yan, K., Larios, E. LeDevedec S.E., van de Water, B., and Verbeek, F.J.(2011), *"Automation in Cytomics: Systematic Solution for Image Analysis and Management in High Throughput Sequences.",* Proceedings IEEE Conf. Engineering and Technology (CET 2011), Vol 7. Shanghai 2011, 195-198

Chapter Summary

In the Chapter 4 and Chapter 5, we have demonstrated two case studies on the HT/HC live cell screen. These two case studies are producing an amount of data on the scale of terabytes, which makes the organization and storage of these image data nontrivial. Additionally, phenotypical measurements of these HT/HC data must be integrated with other omics resources such as Entrez DB, Ensembl DB, BLAST engine to allow the construction of genotype-to-phenotype models. To that end, an HT/HC database management system (HT/HC database system) is required to provide a platform for both data management and integration.

Initially, the HDF5 format, a self-contained data storage file format has been employed in our study as a hierarchical data storage solution. The HDF5-based storage has proven to be an efficient data storage and transportation solution. However, it lacks a build-in searching and querying mechanisms allowing examine data in a comprehensive manner. To overcome the limitation self-contained data storage, we proposed a prototype HT/HC database system. This database system is not only designed for the management of HT/HC image data but also facilitates the integration with omics via a flexible programming API. Together with tools for data mining and semantic analysis tools, the HT/HC database system will eventually provide a multi-layer view of an experiment, which cannot be easily accomplished by self-contained file format such as HDF5.

Several major challenges are encountered during the design of this HT/HC database system. Compared to other designs of concurrent database applications, the design of HT/HC database system faces the following challenges: (1) megabyte-scale data must be locked per transaction, (2) complex image analysis may lead to longer execution duration per transaction, and (3) a multithread-safe programming API is required when accessing data.

In this chapter, a prototype design of HT/HC data management system is proposed to meet with these challenges. The chapter first introduces the principle design of the architecture of the HT/HC data management system. Subsequently, the architecture is further divided into a multilayer model which is comparable to the layered foundation of the OSI model. In the following sections, each layer is explained.

6.1. Principle Design of HT/HC Management System

Initially, we focus on designing a seamless automated framework to complete the data processing for HT/HC screen studies [145][146]. Since the purpose of this system is straightforward, here we employ a task-driven user-centered design principle taking the empirical workflow of the HT/HC screen study (cf. Figure 1-1) as starting point. A HT/HC analysis first starts with the design of the experiment setting based on biological interests. The experiment design is further stored as plate designs. Subsequently, the specimens are visualized using microscopy. The captured images are uploaded to a network-attached storage (NAS) and quantified using image analysis. Eventually, the phenotypical measurements are further probed using machine learning and statistical analysis to provide a comprehensive data representation. From the workflow (cf. Figure 1-1), we defined a simplified functional division between each task during the HT/HC analysis. Moreover, the essential data type passed between each task is also defined. From the definitions, the general design of HT/HC data management system is illustrated as Figure 6-1.

The general design follows the layered architecture similar to the OSI model [147]; as we descend each layer is closer to the raw data. During the design of our HT/HC data management system, we first start from the analysis workflow of an empirical HT/HC experiment (cf. Figure 1-1). Compared to a software design following a requirement-driven approach, our approach is very applicable as there are no clear mutual requests definitions underlying different HT/HC screen experiments. The design of HT/HC screen experiment is often semi-systematic and there are very little similarities in the organization of experiment materials and perturbations employed in the experiment design (cf. Ch. 4 and Ch. 5). Moreover, the existing similarities are mostly related to the raw data produced during HT/HC screen experiment instead of biological materials. Therefore, the analysis workflow of HT/HC empirical experiments is a good starting point to study the limited amount of similarities among HT/HC screen experiments. This workflow illustrates the major steps in a HT/HC screen experiment.

From the HT/HC analysis workflow (cf. Figure 1-1), two element types are defined. (1) The **procedure element** type defines a number of essential steps in a HT/HC experiment. (2) The **data element** type defines all sorts of basic raw data produced and transported in between procedures. Each procedure element receives an input data element and performs a collection of operations to produce an output data element. To that end, each procedure element is considered to be a **functional module** [148][149]. Furthermore, here we further generalized the HT/HC analysis workflow into a top-down **layered architecture** (cf. Figure 6-1). Each layer represents a collection of functional modules in the HT/HC analysis workflow.

From the highest to the lowest layer (cf. Figure 6-1), these layers include (1) **end-user GUI layer**, (2) **WS-API layer**, (3) **web service host layer**, (4) **database layer**, and (5) **computation layer**. The top layer, namely the **end-user GUI layer**, is an encapsulated user interface that allows the end-user to access all functional modules with a general view. From this layer, the end-users are given the opportunity to perform a collection of customized pipelines of HT/HC screen analysis without having to understand the architecture of the functional modules such as image analysis, data analysis, and data management. The seamless integration of all lower layers functional modules is accomplished via the implementation of the **Web Service based**

Application Programming Interface (WS-API) layer. A **Web Service (WS)**, by World Wide Web Consortium (W3C) definition [150], is a software system designed to support interoperable machine-to-machine interaction over a network. It has an interface described in a machine-processable format (specifically Web Services Description Language (WSDL)). Other systems interact with the Web service in a manner prescribed by its description using Simple Object Access Protocol (SOAP) messages, typically conveyed using HTTP with an XML serialization in conjunction with other Web-related standards.

The **WS-API layer** is in fact a pseudo layer in which the functional module is wrapped in a web service shell. Such design takes advantage of the compatibility of WS-API and allows module developed in different programming languages to be able to communicate with each without implementing complex cross-language programming. Moreover, with the web service shell, updating in each module does not necessarily require additional changes in the dependent module. To physically host the WS-API layer, the **web service host layer** is introduced. These web service server hosts the foundations of programming API and software architecture. The **image database layer** is the major management layer providing an organized storage and distribution of image data, auxiliary image results, phenotypical measurements and supplementary documentation. The lowest layer, namely the **computation layer**, is the foundation layer which provides computational power for all higher layers.

Following the top-down design of each layer, this chapter is organized as follows. In the next sections, we will focus on end-user GUI layer, WS-API layer, and the database layer since the computation layer is beyond the scope of management and the web service host layer employs only commercial and open-source software. Finally, we will address the database layer.

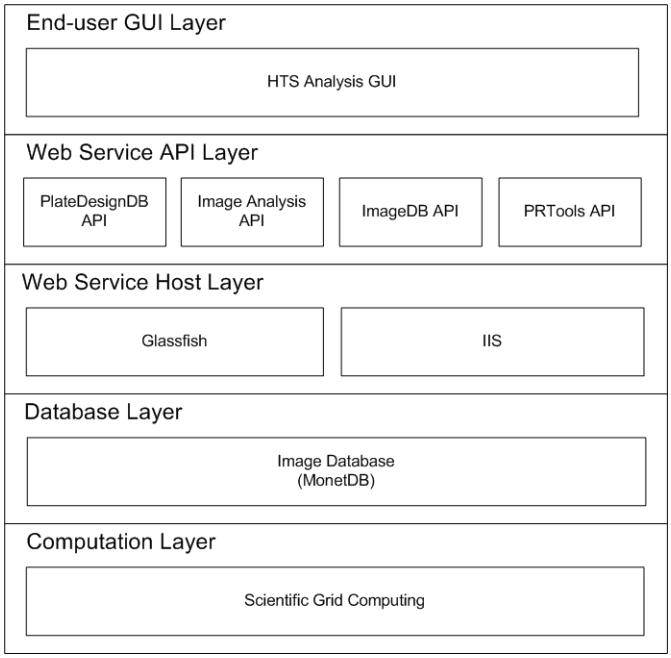


Figure 6-1 the structure of HTS analysis platform

6.2. Implementation of Layers

This section illustrates the fundamental design of the system architecture. The section follows the top-down design flow of the layered model depicted in Figure 6-1. In this section, the top three layers will be described. We will first start with the highest layer.

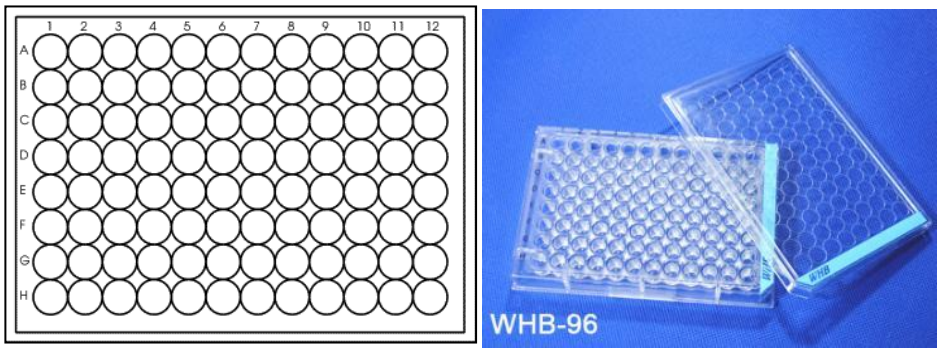
6.2.1. End-user GUI Layer

The end-user GUI layer is the highest layer providing a data input/output and a visualization mechanism. It is designed to be both starting point and end point of a HT/HC analysis. This layer is originally designed as a solution for metadata management in HT/HC such as the bookkeeping of experiment protocols. As introduced in Figure 1-1, HT/HC screen study starts with the design of the experiment; during which the researcher decides which materials and what conditions will be included. Such design is often referred as **plate design**.

The **plate design** begins with the researcher first chooses one cell culture plate as the design model. A cell culture plate (cf. Figure 6-2a) is a plastic or glass plate containing a number of small wells each resembling an individual Petri disk that can hold live specimens. The actual dimensionality, meaning width and length, of cell culture plate is fixed. The numbers of rows and columns of the wells within the culture plate can be in 4x6, 6x8, 8x12 or even higher. A higher number of well allows more experimental perturbations (conditions) to be considered in one experiment. However, more wells also mean longer imaging time during HT/HC screen since, for microscopy, there always is a minimum duration to the imaging of each well. Depending on the complexity of experiment, the researcher often chooses a culture plate with sufficient wells to hold most experimental conditions meanwhile it should guarantee the shortest imaging duration as possible. To visualize the design procedure, here we introduce two empirical spreadsheet-based plate designs of HT/HC screen.

Figure 6-2b is the experimental design of the growth factor regulation experiment described in Ch. 4. In this study, the researcher wants to extract phenotypical variability of cancer cells under the influence of different growth factors. To achieve this purpose, the researcher first selected four growth factors and two cell types from one cancer cell line. Each cell type will be treated with a single or combination of growth factors. Therefore, the researcher divides one 8x12 culture plate into left and right section. The designated growth factors are mapped into the row and column of and stored in a spreadsheet.

Figure 6-2c is the experiment design of a dynamic study EGF transportation pattern [29]. In this study, the researcher wants to quantify the transportation model of EGF in cells under different treatments. Again, two cell types are chosen for the study. Instead of employing live cell imaging, the researcher decided to fix cells at each time point including 10 min, 20 min, 40 min, 80min, and 160 min (cf. Figure 6-2c [column]). Similarly, the whole culture plate is divided into left and right section to house two cell types and each column represent the fixing time of cells. The row represents different treatment the cells received.



(a) a 8x12 cell culture plate

DMSO	HGF	EGF+HGF	FGF+TGFbeta	DMSO	HGF	EGF+HGF	FGF+TGFbeta
DMSO	HGF	EGF+HGF	FGF+TGFbeta	DMSO	HGF	EGF+HGF	FGF+TGFbeta
EGF	TGFbeta	EGF+TGFbeta	HGF+TGFbeta	EGF	TGFbeta	EGF+TGFbeta	HGF+TGFbeta
EGF	TGFbeta	EGF+TGFbeta	HGF+TGFbeta	EGF	TGFbeta	EGF+TGFbeta	HGF+TGFbeta
FGF	EGF+FGF	FGF+HGF	all	FGF	EGF+FGF	FGF+HGF	all
FGF	EGF+FGF	FGF+HGF	all	FGF	EGF+FGF	FGF+HGF	all
plGFP				GB1			

(b) Spreadsheet-based data management: growth factor regulation

	plGFP					GB1				
	10 min	20 min	40 min	80 min	160 min	10 min	20 min	40 min	80 min	160 min
ctrl										
mock										
+EGF										

(c) Spreadsheet-based data management: EGF translocation experiment

Figure 6-2 plate design in HT/HC screen

When conducting a wet-lab experiment, this plate design will be followed. During the image analysis and data analysis, the same spreadsheet will also be used as bookkeeping information for data comparison. However, for large scale experiment such as siRNA functional screens, hundreds siRNA targets (conditions) will be manually mapped into the plate designs and each experiment requires 20~40 different plate designs to hold all targets. It is difficult to produce an error-free scenario when performing the design. Moreover, an error in the plate design will lead to a false-conclusion during image and data analysis since the original expectation of output may no longer fulfill due to the mismatch between bookkeeping information and the practical experiment being conducted. Therefore, to minimize or eliminate the manual process of the important bookkeeping mechanism, namely the plate design, here we design a seamless GUI for the plate design in HT/HC screen studies.

The plate design GUI (cf. Figure 6-4) starts with an review of the design of a plate[145]. From a number of empirical studies, we extract the basic elements employed in most HT/HC screen studies. We introduce the concept of **plate**, **well**, **condition**, and **group** (cf. Figure 6-3). The design concept is to emulate the natural plate design procedure that researchers have used. To that point, the design of the plate design GUI follows a top-down flow instead of a requirement-driven. The **plate** is a representation of plate layout. Each plate contains a number of wells. There is fixed a one-to-many relationships between plate and wells. Instead of giving a fixed row and column number, we allow users to choose the number-id of row and column and fit a plate layout with the number of wells. The **well** represents a physical well in a culture plate. Each well contains a number of **conditions** which represents the actual treatment, treatment duration, cell type or any experiment perturbations that may be employed in the HT/HC screen. There is a many-to-many relationship between conditions and

wells since a well can have a combination of perturbations (conditions). The **group** is a higher level concept of wells following some reasoning. For example, Figure 6-2c contains two groups which represent different cell types.

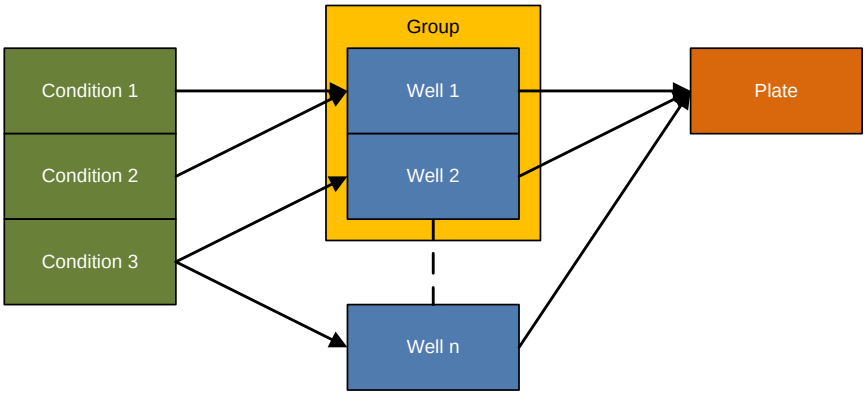


Figure 6-3 Components in plate design

With the GUI design (cf. Figure 6-4), we allow the end-user to first create an empty plate of chosen size. Then the end-user imports the master sheet of conditions which contains all experiment perturbations; meanwhile conditions can be added to the wells. Image analysis of logical groupings can be then performed. The major improvement of this GUI design over spreadsheet-based design is the introduction of what-you-see-is-what-you-get (**WYSIWYG**) principle and **drag-&-drop** design.

The **WYSIWYG** design principle provides the foundation for GUI of plate design. Instead of using a table to emulate a plate layout, here we employ a row-column well map layout which is an exact map of the physical plate. It provides a straightforward overview of the whole plate without losing the possibility of zooming into a single-well.

The **drag-&-drop** design allows a faster mapping from the master list of conditions to desirable well in the plate. By allowing multiple selections, the same condition can also be assigned into multiple wells with a single dragging; similarly, one well can be assigned with multiple conditions with a single dragging. Moreover, with colored labels for conditions, wells contain different conditions can be easily distinguished by color (cf. Figure 6-4). When displaying different conditions together, colors can be combined or switched off to improve fast recognition. In the screenshots (cf. cf. Figure 6-2c), the similar operation using spreadsheet-based solution would require end-users to constantly switch between spreadsheets. Moreover, wells containing combined conditions cannot be easily visualized together.

In Figure 6-6, the sequence diagram of the use of a plate design GUI is displayed. The user first finishes plate designs using the plate layout design interface (cf. Figure 6-4) and based on the design user will conduct wet-lab experiment. After collecting all images associated with the plate design, the user can upload all images and map them to the database using the ImageDB API. Subsequently, the user will initialize analysis of the screen and then the GUI will invoke image analysis API and pattern recognition API to produce intermediate results of the image and data analysis (cf. Figure 6-5a & b). By using different data analysis approaches, the user

may verify the initial hypothesis or compare phenotypical characteristics under different treatments.

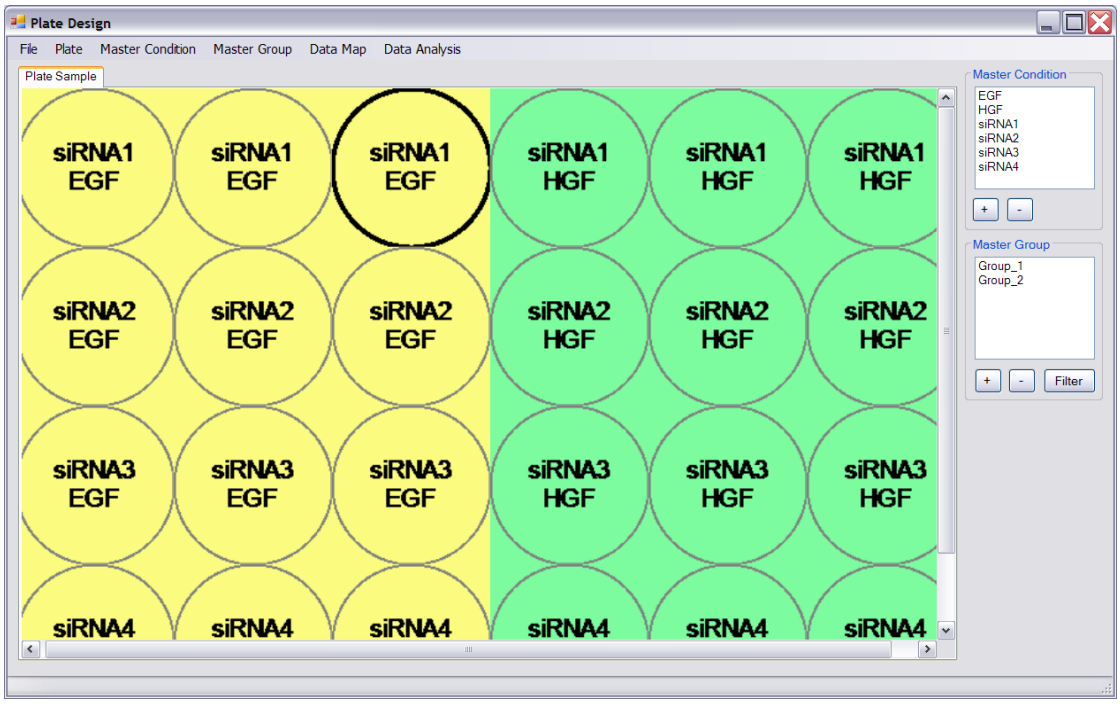
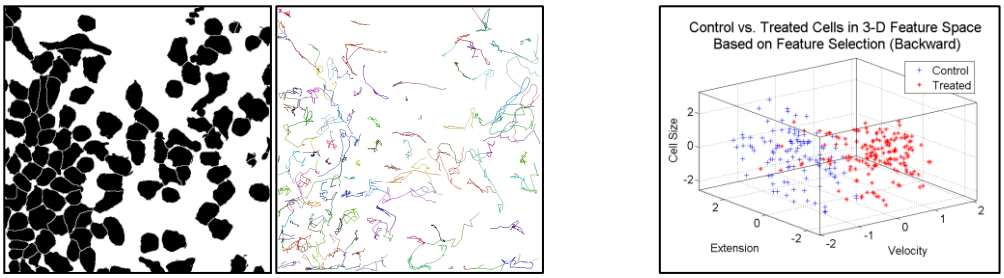


Figure 6-4 plate design GUI



(a) intermediate results of image analysis **(b) data analysis result**

Figure 6-5 front end of the data management system

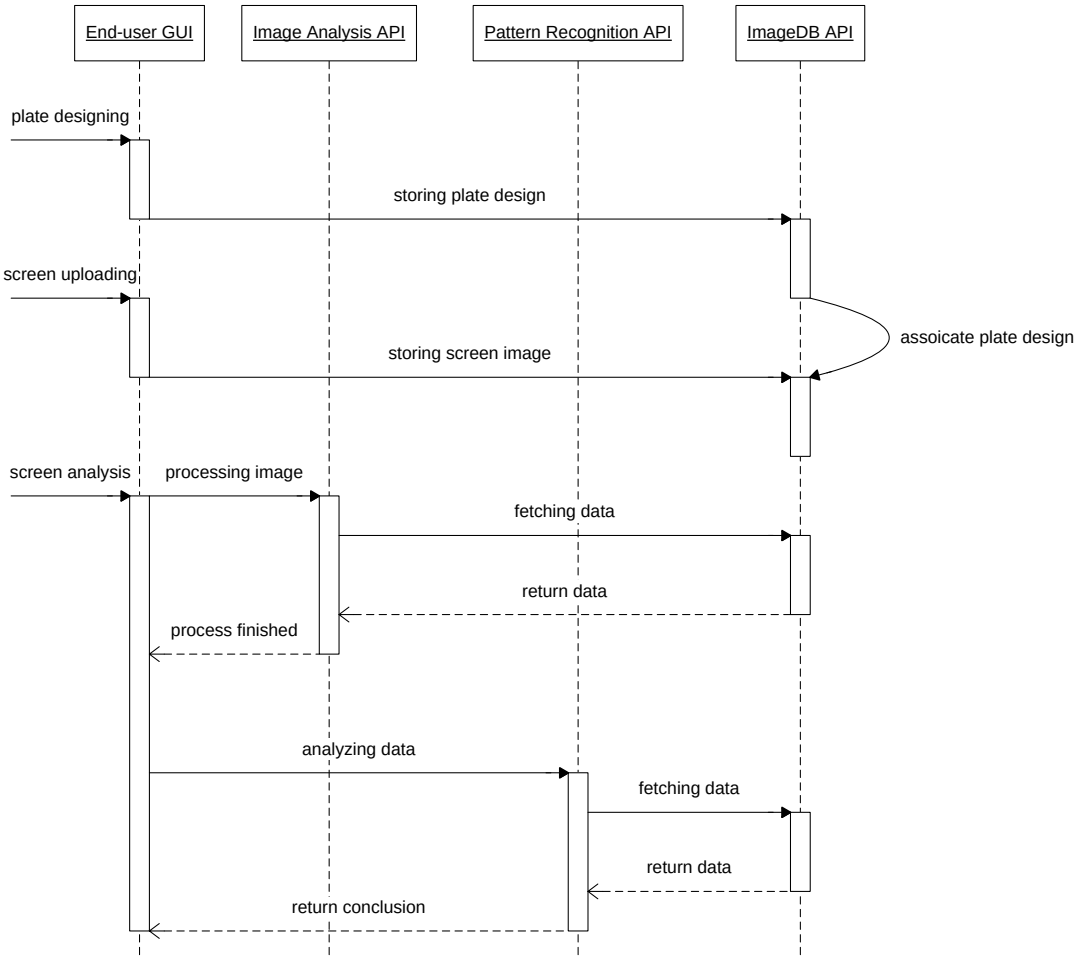


Figure 6-6 sequence activity diagram of using end-user GUI

6.2.2. WS-API Layer

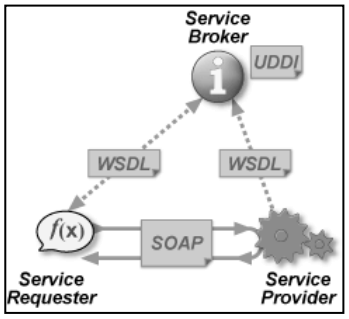


Figure 6-7 web service architecture

In order allow different functional modules to communicate with each other, a universal wrapping application programming interface (API) is required. Although it is possible to hardcode cross-module communication, such implementation may limit the extensibility of software and violates the encapsulation design. Therefore, we adapted the principle of **Web Service based Application Programming Interface (WS-API)** as the solution. The W3C organization defines a "Web service" (WS) as "a software system designed to support interoperable machine-to-machine interaction over a network" (cf. Figure 6-7) [150].

To take advantage of the flexibility of the Web Service, the WS-API of each function module is wrapped in a Web Service shell by following modular programming and published by web server supporting Web Service, such as Java Glassfish and Microsoft IIS. We will first use the ImageDB API as a sample to illustrate the abstract design of WS-API communication (cf. Figure

6-8). Furthermore, we will briefly introduce the implementation of other functional modules such as plate design API, image analysis API, and pattern recognition API.

ImageDB API

The ImageDB API is a WS-API that grants database accessibility to other function modules. The design of ImageDB WS-API is illustrated in Figure 6-8. Such design provides several advantages:

1) It is possible to access data set without understanding the complex internal relationships
2) Internal structure or physical location of data are not visible to the end-users
3) Internal modifications do not require updates at client state
4) The modular design significantly improves extensibility and flexibility of cross-module communications.

With the ImageDB API, other modules may exchange data without direct communication with each other.

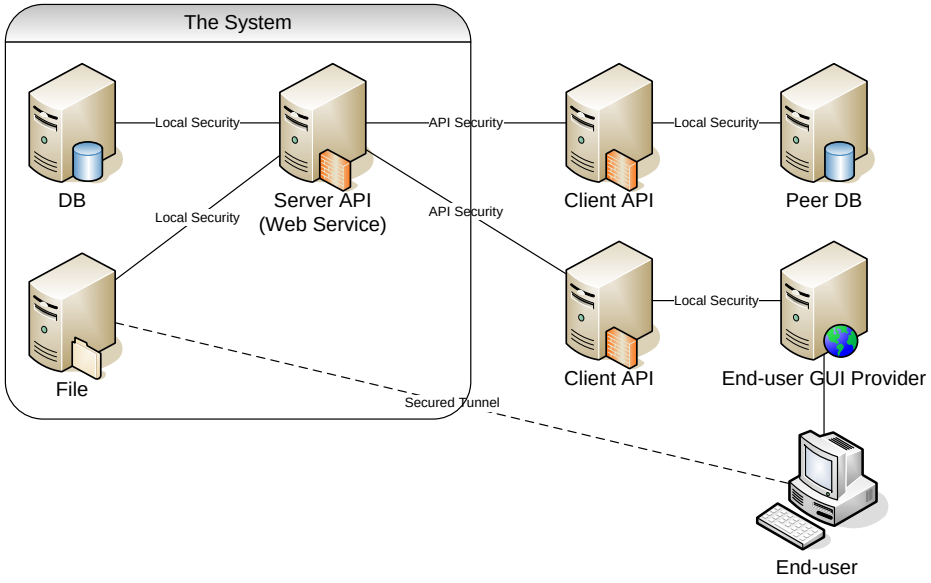


Figure 6-8 abstract design of WS-API communication

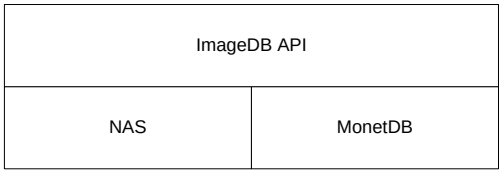


Figure 6-9 underline structure of the ImageDB API

In the design of ImageDB API, we have chosen for the MonetDB DBMS [151][152]. MonetDB is a leading open source database system that has been dedicated to the management of large datasets. Compared to other DBMSs, it is well-known for its performance in processing analytical queries on large scale data sets. Instead of storing data in a row-based memory block, the MonetDB stores data in a column-based memory block (cf. Figure 6-10). Thus, when processing analytical queries, only the required columns are loaded. Such design is in particular important when data contain extensive amounts of primary keys or if not all column data will

be involved in the analytical queries. According to our study [145][146], such storage structure provides an increased access speed since the conditional filtering can be executed much faster.

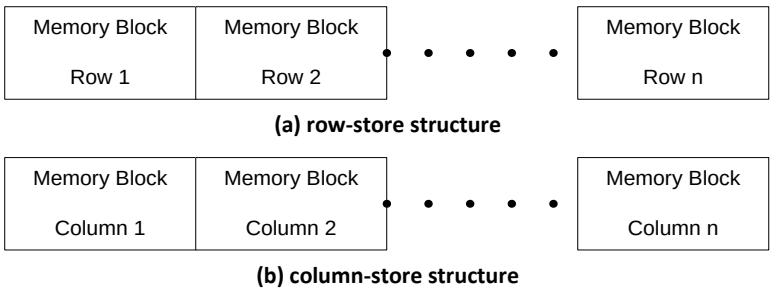


Figure 6-10 row-store structure vs. column-store structure

Plate Design Module

The plate layout design provides a graphical user interface allowing end-users to rapidly deploy, modify, and search through plate designs, to which auxiliary data such as experimental protocols, images, analysis result and supplementary literature is attached. In addition the plate design provides fast cross-reference mechanism in comparing data from various origins. This module is also used as the front end for the visualization of results such as using heatmaps, cell detection, or motion trajectories. The underlying structure of the plate design module is illustrated in Figure 6-11.

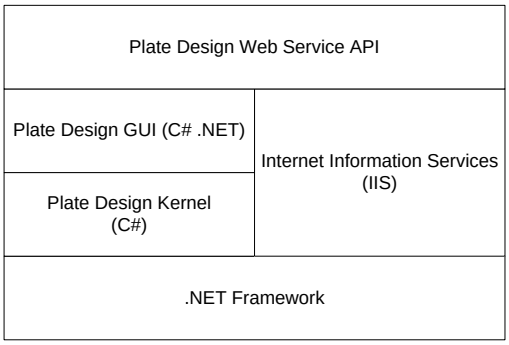


Figure 6-11 underline structure of the plate design API

The plate design module is constructed in the JAVA language. Since most of the end-users are more familiar with Windows-based applications such as Excel, the Windows-based GUI using JAVA may further improve user satisfaction. Compared to other programming languages, JAVA is a platform independent programming language which allows the GUI to be directly run on computers with a Java Virtual Machine (JVM) installed without the need to recompile the code.

Image Analysis Module

The customized image processing and analysis is applied to obtain phenotypical measurements for each of the different treatments. An open source image processing kernel, i.e. ImageJ, is extended with packages providing customized and robust image segmentation and object tracking algorithms dedicated to various types of cytomics (cf. Figure 6-5b). The current package covers solutions to cell migration, cellular matrix dynamics and structure dynamics analysis (cf. Chapter 4 & Chapter 5). It has been practiced in HTS experiments for toxic compound screening of cancer metastasis [10][15], wound-and-recovery of kidney cells [70] and cell matrix adhesion complex signaling[29], etc. This module (cf. Figure 6-12) is designed as

WS-API. As the image analysis computation requires large image volumes to be processed, GRID computing is used to obtain results in reasonable time.

The current WS-API serves as a wrapping class around the image analysis package kernel and transforms the functional module into a black box that can be access by other functional module.

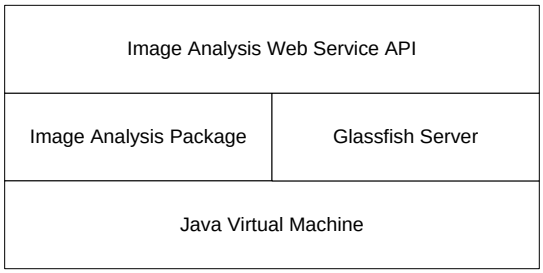


Figure 6-12 underline structure of the image analysis API

Data Analysis Module

To maximize the availability of data analysis tools, we have included the PRTools package [153] as the machine-learning toolset kernel for the data analysis module. Furthermore, we develop a customized pattern recognition toolset for data analysis with spatio-temporal data. The data analysis module (cf. Figure 6-13) is implemented as WS-API using .NET output of MATLAB deployment tools. Such architecture allows a rapid adaptation to complex mathematical algorithms. Furthermore, the flexibility of GUI-based data mining procedures can be operated by the end-user with a minimum of knowledge on machine learning.

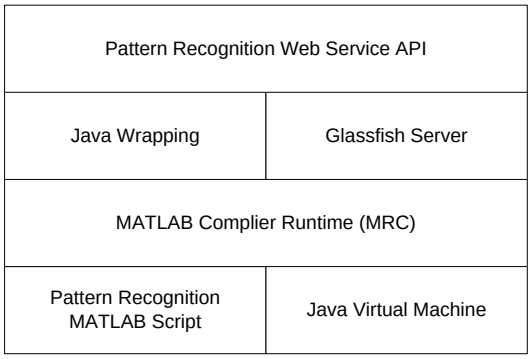


Figure 6-13 underline structure of the pattern recognition API

Since we implemented centralized and platform-independent software architecture, end-users can update algorithms in real-time without compatibility or package dependency.

6.3. Conclusion and Discussion

The beta test shows that a workload previously taking one month can now be accomplished within a week using the HT/HC database system. The major workload is now mostly on experiment preparation and image acquisition since these two procedures are labor-intensive. Normally, the experiment preparation takes one day to accomplish while image acquisition will take another day. However, the image and data analysis of this image set takes a couple of hours. Complex experiment design requires more time to prepare and often prone to errors. In contrast, image analysis and data analysis can easily adapt the increase of in the amount of data by simply adding more computational power.

In terms of software design, the HT/HC DMS follows a modular design and all modules are implemented in the form of web services, therefore, updating the system is virtually instantaneous. Moreover, this framework is very flexible as it allows connecting other web services. Consequently, a fast response to new progress in image and data analysis algorithms can be realized. Additionally, the seamless design of HT/HC database system has significantly decreases manual errors in the data transportation, image analysis and data analysis. Comparing with solutions such as CellProfiler[82] or ImagePro, our solution provides a unique approach for HT/HC image analysis. It allows end-users to perform high-profile HT/HC analysis with a minimum level of prior experience on image analysis and machine learning. Moreover, the modular design allows faster connecting with other web services such as BLAST [154]. Consequently, a faster response to progress in image and data analysis can be realized. Further integration with online bio-ontology databases and open gene-banks is considered so as to allow integration of the data from other sources. Therefore, the platform can eventually evolve into an interdisciplinary platform for cytomics.

Acknowledgement

The authors would like to thank Wouter Zomervrucht, Alice Bodanzky, Tijl Kindt, Arthur Stuivenberg, and Leah Winkel for their meticulous work on developing the prototype of the HT/HC data management system.

Chapter 7

Conclusion and Discussion

7.1. Conclusions

This thesis explored solutions for image segmentation and object tracking for high-throughput/high-content (HT/HC) screens in cytomics studies. To demonstrate the usefulness of the solution, two case studies of cancer migration were discussed to elaborate the efficiency and limitations of several generic image and data analysis approaches frequently observed in HT/HC screen studies. From an analysis of these limitations, a number of dedicated algorithms are developed and evaluated. The evaluations show that the dedicated algorithms, namely watershed masked clustering (WMC) segmentation [62][29][15][10], kernel density estimation (KDE) with mean shift tracking algorithm [15][10][9], and energy-driven linear (EDL) model tracking algorithm, provide a more robust and accurate performance. From these dedicated algorithms, our case studies further demonstrate the possibility to produce an objective understanding of each unique phenotypic characteristic using morphology and motility measurements from both cellular and subcellular level.

With a good automated data management in place [145][146], the analysis pipeline can be further performed in a more efficient manner. The data management system shields end-users from the underlying complexity of the computational approaches in the pipeline by providing an integrated high-end GUI. The automation beneath the GUI will enable to scale to a higher volume of image data which is custom to HT/HC screen, i.e. terabyte level. It may further increase the data accessibility and interdisciplinary data conformity by standardizing different HT/HC image formats and measurement structures. The interdisciplinary data conformity is an important feature since the ultimate goal of cytomics and quantitative microscopy is to integrate all –omics data to provide a numeric modeling of genotype-to-phenotype mechanism.

Chapter 2 Robust Image Segmentation for Cytomics

The Chapter 2 has illustrated a dedicated segmentation algorithm, namely **Watershed Masked Clustering (WMC) algorithm**, for high-throughput cytomic studies. When compared to other segmentation algorithms, the WMC is capable of producing an accurate segmentation results when image contains objects with nonlinear intensity and morphology variation. Such a trait is particularly useful in HT/HC screen studies since the responses of treatments are often unclear prior to the experiment.

Chapter 3 Robust Object Tracking for Cytomics

The Chapter 3 has illustrated two dedicated object tracking algorithms, namely the **Kernel Density Estimation (KDE) with Mean Shift tracking algorithm** and the **Energy Driven Linear (EDL) model tracking algorithm**. Compared to other tracking algorithms, these two algorithms show a robust tracking performance for cellular and subcellular objects without manual intervention. Moreover, they can be extended to other application domains if new motion models can be built.

Chapter 4 A Study to Cell Migration Analysis

In this case study, we have numerically extracted the morphology and motility characteristics of random cancer migration under the influence of different growth factor treatments using image and data analysis solutions described in Chapter 2 and Chapter 3. Compared to manual analysis, it is clear that an automated image and data analysis solution can provide a more

objective and reproducible understanding of a cell biology experiment. It is an important trait in cytomics studies since it allows the experiment to scale to a larger volume of data without scarifying the quality of analysis while it is impossible with manual analysis.

Chapter 5 A Study to Dynamic Matrix Adhesion Analysis

In this case study, we have demonstrated an integrated solution to quantify the morphology and dynamics of both matrix adhesions and cells using the automated image analysis solution described in Chapter 2 and Chapter 3. From the measurements, we further confirm several correlations between matrix adhesion dynamics and cell migration. This case study [16] is one of the few attempting to reveal the subcellular control mechanism behind random cell migration using an automated method.

Chapter 6 Reasoning over Data in HT/HC Management

In Chapter 6, we have demonstrated the design and implementation of a dedicated data management system. In our study (Ch. 4 & 5) the data management system serves as both the starting and the end point of the analysis by shielding the end-user from underlying complexity with a high-end integrated GUI [146][145]. Thus, it allows end-user to perform sophisticated HT/HC screen studies without an extended knowledge of image and data analysis.

7.2. Discussion

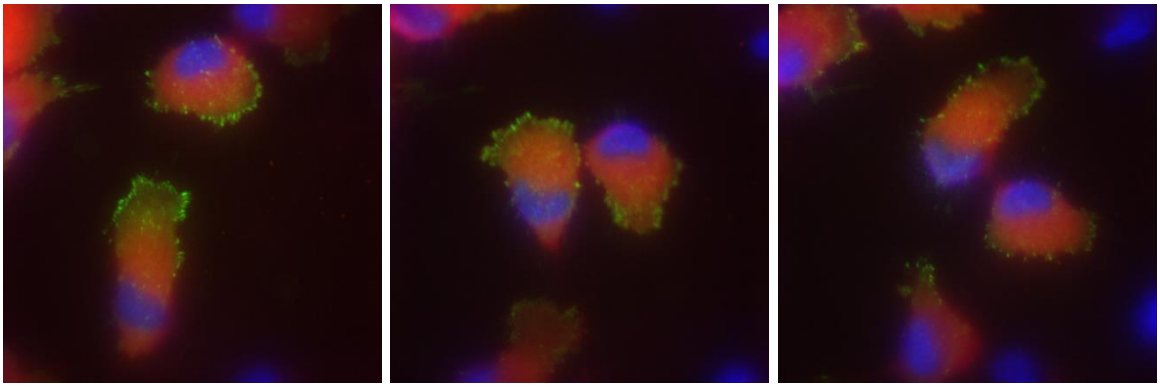
This thesis has explored the image and data analysis solutions for empirical HT/HC screen study. In the future, we will focus on making improvements in the following aspects.

1. Experiment Preparation in HT/HC Screening

Although, experiment preparation is beyond the scope of this thesis, in the two case studies (cf. Chapter 4 and Chapter 5) it has demonstrated how image quality is affected by the experiment preparation. At current stage, the cell-to-cell variability [19] is a typical challenge frequently encountered in our research. Often cell behavior is subjected to number of factors such as cell age, local cell density, individual mutation, etc. It is possible to employ a normalization strategy similar to microarray data analysis [155][156][157], but it would reveal more information if the true mechanism behind the cell-to-cell variability can be understood. Dedicated research should be conducted to gain understanding of these phenomena.

2. Image Acquisition

Modern microscopes are frequently equipped with additional functional controllers allowing an on-the-fly adaptation of the microscope settings such as focus and position of specimens. However, these adaptation mechanisms are subjected to the heuristics from which the adaptation algorithm is built. For example, in the experiments underlying the work in Chapter 5, frequently out-of-focus errors were encountered (cf. Figure 7-1). In addition, in live cell migration or subcellular dynamics study the temporal-resolution of image sequences is often sacrificed in exchange of a better image quality. Both issues are hardware related that can be significantly improved by new developments in the microscope technology.



Frame 1 focused | Frame 50 out of focus | Frame 100 slightly out of focus

Figure 7-1 out-of-focus error during live cell imaging

3. Image Analysis

Tuning of Segmentation

We believe that a mechanism of robust self-adaptation is the key to a successful image segmentation solution in HT/HC screen studies. For this adaptation mechanism, a robust representation of image heuristics [62][62] is crucial for its success. The sophistication of the self-adaptation mechanism can benefit most from new progress in machine vision [158].

Phenotypic Measurement

The phenotypic measurements used in this thesis are capable of extracting each unique phenotypic profile of objects. However, they are often not scale-free [33][32][31] and cannot be compared across different microscopy settings. Moreover, for some cell phenotype, the current measurement is not optimized. For example, the velocity is often measured from the difference between objects in consecutive frames, but the difference is not necessarily only associated with motility (cf. Figure 7-2) [140][159]. A local regression approach [126] is frequently used to extract the majority trend of migration while a more sophisticated motion modeling solution [140][159] is preferred for the measuring of motility.

(a) shifting of mass center is correlated to motion **(b) shifting of mass center is correlated to deformation**

Figure 7-2 a Z-projection plot of cell body contours during migration

4. Data Analysis

In the case studies of Chapter 4 and Chapter 5, we use low-order statistics to provide some characterizations of cellular and subcellular dynamics; previously unknown or unverified. However, by losing the temporal dimension, it is very easy to overlook particular dynamic behavior behind control mechanism of cell migration. Therefore, the study of expanding temporal statistical analysis into temporal-spatial data is required.

The HT/HC screen study is an emerging field in cytomics research. Together with image and data analysis it provides an efficient analysis tool for studies in functional genomics and cell

biology. There are challenges to be met in image and data analysis of HT/HC screen studies. However, with new approaches developed, it will be eventually possible to accomplish the genotype-to-phenotype modeling. This requires combining data from different aspects of the same study-object. Cytomics can provide accurate and well annotated data for such boarder view.

References

[1] P. Baldi and S. Brunak, *Bioinformatics: The Machine Learning Approach*, 2nd ed. MIT Press, 2001.

[2] B. Hesper and P. Hogeweg, "Bioinformatica: een werkconcept," *Kameleon*, vol. 1, no. 6, pp. 28–29, 1970.

[3] *Oxford English Dictionary*. 2013.

[4] U. Maskos and E. Southern, "Oligonucleotide hybridizations on glass supports: a novel linker for oligonucleotide synthesis and hybridization properties of oligonucleotides synthesised in situ," *Nucleic Acids Res*, vol. 20, no. 7, pp. 1679–84, 1992.

[5] C. B. Black, T. D. Duensing, L. S. Trinkle, and R. T. Dunlay, "Cell-based screening using high-throughput flow cytometry.," *Assay Drug Dev. Technol.*, vol. 9, no. 1, pp. 13–20, 2011.

[6] W. H. De Vos, L. Van Neste, B. Dieriks, G. H. Joss, and P. Van Oostveldt, "High content image cytometry in the context of subnuclear organization.," *Cytometry. A*, vol. 77, no. 1, pp. 64–75, Jan. 2010.

[7] M.-A. Bray, A. N. Fraser, T. P. Hasaka, and A. E. Carpenter, "Workflow and metrics for image quality control in large-scale high-content screens.," *J. Biomol. Screen.*, vol. 17, no. 2, pp. 266–74, Feb. 2012.

[8] A. Look and S. Melvin, "Aneuploidy and percentage of S-phase cells determined by flow cytometry correlate with cell phenotype in childhood acute leukemia," …, vol. 60, no. 4, pp. 959–67, Oct. 1982.

[9] K. Yan, S. LeDévédec, B. van De Water, and F. Verbeek, "Cell Tracking and Data Analysis of in vitro Tumour Cells from Time-Lapse Image Sequences," in *VISAPP 2009*, 2009, pp. 281–287.

[10] S. LeDévédec, K. Yan, H. de Bont, V. Ghotra, H. Truong, E. Danen, F. J. Verbeek, and B. vande Water, "A Systems Microscopy Approach to Understand Cancer Cell Migration and Metastasis," *Cell. Mol. Life Sci.*, vol. 67, no. 19, pp. 3219–3240, 2010.

[11] K. J. Simpson, L. M. Selfors, J. Bui, A. Reynolds, D. Leake, A. Khvorova, and J. S. Brugge, "Identification of genes that regulate epithelial cell migration using an siRNA screening approach.," *Nat. Cell Biol.*, vol. 10, no. 9, pp. 1027–1038, 2008.

[12] B. Neumann, T. Walter, J.-K. Hériché, J. Bulkescher, H. Erfle, C. Conrad, P. Rogers, I. Poser, M. Held, U. Liebel, C. Cetin, F. Sieckmann, G. Pau, R. Kabbe, A. Wünsche, V. Satagopam, M. H. A. Schmitz, C. Chapuis, D. W. Gerlich, R. Schneider, R. Eils, W. Huber, J.-M. Peters, A. A. Hyman, R. Durbin, R. Pepperkok, and J. Ellenberg, "Phenotypic profiling of the human genome by time-lapse microscopy reveals cell division genes.," *Nature*, vol. 464, no. 7289, pp. 721–727, 2010.

[13] J. De Rooij, A. Kerstens, G. Danuser, M. A. Schwartz, and C. M. Waterman-Storer, "Integrin-dependent actomyosin contraction regulates epithelial cell scattering.," *J. Cell Biol.*, vol. 171, no. 1, pp. 153–64, Oct. 2005.

[14] R. Pepperkok and J. Ellenberg, "High-throughput fluorescence microscopy for systems biology," *Nat. Rev. Mol. Cell Biol.*, vol. 7, no. 9, pp. 690–696, 2006.

[15] L. Damiano, S. E. Le Dévédec, P. Di Stefano, D. Repetto, R. Lalai, H. Truong, J. L. Xiong, E. H. Danen, K. Yan, F. J. Verbeek, E. De Luca, F. Attanasio, R. Buccione, E. Turco, B. van de Water, and P. Defilippi, "p140Cap suppresses the invasive properties of highly metastatic MTLn3-EGFR cells via impaired cortactin phosphorylation," *Oncogene*, vol. 30, no. 5, pp. 624–33, 2011.

[16] K. Yan, S. Le Dévédec, B. van de Water, and F. J. Verbeek, "AUTOMATED ANALYSIS OF MATRIX ADHESION DYNAMICS IN MIGRATING TUMOR CELLS."

[17] B. Neumann, M. Held, U. Liebel, H. Erfle, P. Rogers, R. Pepperkok, and J. Ellenberg, "High-throughput RNAi screening by time-lapse imaging of live human cells," *Nat. Methods*, vol. 3, no. 5, pp. 385–390, 2006.

[18] R. D. Goldman, J. Swedlow, and D. L. Spector, *Live Cell Imaging: A Laboratory Manual*. CHS-Press, 2005.

[19] B. Snijder and L. Pelkmans, “Origins of regulated cell-to-cell variability.,” *Nat. Rev. Mol. Cell Biol.*, vol. 12, no. 2, pp. 119–125, Jan. 2011.

[20] I. C. Ghiran, “Introduction to fluorescence microscopy.,” *Methods Mol. Biol. Clift. Nj*, vol. 689, no. 1, pp. 93–136, 2011.

[21] J. W. Lichtman and J.-A. Conchello, “Fluorescence microscopy.,” *Nat. Methods*, vol. 2, no. 12, pp. 910–919, 2005.

[22] A. Hoffman, M. Goetz, M. Vieth, P. R. Galle, M. F. Neurath, and R. Kiesslich, “Confocal laser endomicroscopy: technical status and current indications.,” *Endoscopy*, vol. 38, no. 12, pp. 1275–1283, 2006.

[23] N. S. Claxton, T. J. Fellers, and M. W. Davidson, “LASER SCANNING CONFOCAL MICROSCOPY,” *Microscopy*, vol. 1979, no. 21, pp. 99–111, 1979.

[24] E. J. Ambrose, “A surface contact microscope for the study of cell movements,” *Nature*, vol. 178, no. 4543, p. 24, 1956.

[25] D. Axelrod, “Cell-substrate contacts illuminated by total internal reflection fluorescence,” *J. Cell Biol.*, vol. 89, no. 1, pp. 141–145, 1981.

[26] R. A. Hoebe, C. J. F. Van Noorden, and E. M. M. Manders, “Noise effects and filtering in controlled light exposure microscopy.,” *J. Microsc.*, vol. 240, no. 3, pp. 197–206, 2010.

[27] J.-Y. Peng, C.-C. Lin, and C.-N. Hsu, *Adaptive Image Enhancement for Fluorescence Microscopy*. IEEE, 2010, pp. 9–16.

[28] G. M. P. Van Kempen, “Image Restoration in Fluorescence Microscopy,” Delft University Press, 1999.

[29] L. Cao, K. Yan, L. Winkel, M. de Graauw, and F. Verbeek, “Pattern Recognition in High-Content Cytomics Screens for Target Discovery - Case Studies in Endocytosis,” in *Lecture Notes in Computer Science*, 2011, pp. 330–342.

[30] A. Pinidiyaarachchi and C. Wahlby, “Seeded watersheds for combined segmentation and tracking of cells,” *Lect. Notes Comput. Sci.*, vol. 3617, pp. 336–343, 2005.

[31] E. Meijering, O. Dzyubachyk, and I. Smal, *Methods for cell and particle tracking.*, 1st ed., vol. 504. Elsevier Inc., 2012, pp. 183–200.

[32] N. Harder, R. Batra, S. Gogolin, N. Diessl, and R. Eils, “Large-SCALE TRACKING FOR CELL MIGRATION AND PROLIFERATION ANALYSIS AND EXPERIMENTAL OPTIMIZATION OF HIGH-THROUGHPUT SCREENS,” *miaab.org*.

[33] M. Held, M. H. A. Schmitz, B. Fisher, T. Walter, B. Neumann, M. H. Olma, M. Peter, J. Ellenberg, and D. W. Gerlich, “CellCognition: time-resolved phenotype annotation in high-throughput live cell imaging,” *Nat. Methods*, vol. 7, no. 9, p. 747, 2010.

[34] R. C. Gonzalez, R. E. Woods, and S. L. Eddins, *Digital Image Processing using MATLAB*. Person Prentice Hall Publishing, 2004.

[35] R. K. Spring, J. T. Fellers, and W. M. Davidson, “Signal-to-Noise Considerations,” *Olympus Microscopy Resource Center*, 2012. [Online]. Available: http://www.olympusmicro.com/primer/techniques/confocal/signaltonoise.html.

[36] T. Würflinger, I. Gamper, T. Aach, and a S. Sechi, “Automated segmentation and tracking for large-scale analysis of focal adhesion dynamics.,” *J. Microsc.*, vol. 241, no. 1, pp. 37–53, Jan. 2011.

[37] R. Padmanabha, L. Cook, and J. Gill, “HTS quality control and data analysis: a process to maximize information from a high-throughput screen.,” *Comb. Chem. high throughput Screen.*, vol. 8, no. 6, pp. 521–527, 2005.

[38] N. Malo, J. A. Hanley, G. Carlile, J. Liu, J. Pelletier, D. Thomas, and R. Nadon, “Experimental design and statistical methods for improved hit detection in high-throughput screening.,” *J. Biomol. Screen. Off. J. Soc. Biomol. Screen.*, vol. 15, no. 8, pp. 990–1000, 2010.

[39] H. Attias and C. E. Schreiner, “Temporal low-order statistics of natural sounds,” *Adv. neural Inf. Process. …*, vol. 9, 1997.

[40] A. a Hill, P. LaPan, Y. Li, and S. Haney, “Impact of image segmentation on high-content screening data quality for SK-BR-3 cells.,” *BMC Bioinformatics*, vol. 8, p. 340, Jan. 2007.

[41] M. Baatz, N. Arini, and G. Binnig, “Object-oriented image analysis for high content screening: Detailed quantification of cells and sub cellular structures with the Cellenger software,” *Cytom. Part A*, vol. 658, no. May, pp. 652–658, 2006.

[42] D. L. Pham, C. Xu, and J. L. Prince, “Current Methods in Medical Image Segmentation,” *Annu. Rev. Biomed. Eng.*, vol. 2, pp. 315–337, 2000.

[43] E. Hancock and J. Kittler, “Adaptive estimation of hysteresis thresholds,” in *Proceedings of Computer Vision and Pattern Recognition*, 1991, pp. 196–201.

[44] A.-P. Condurache and T. Aach, “Vessel segmentation in angiograms using hysteresis thresholding,” in *IAPR Conference on Machine Vision*, 2005.

[45] J. A. Sethian, *Level Set Methods and Fast Marching Methods*, 2nd ed. Cambridge University Press, 1999.

[46] N. B. Venkateswarlua and P. S. V. S. K. Raju, “Fast isodata clustering algorithms,” *Pattern Recognit.*, vol. 25, no. 3, pp. 335–342, 1992.

[47] M. Mancas and B. Gosselin, “Segmentation using a region-growing thresholding,” in *Image Processing: Algorithms and Systems IV*, 2005.

[48] O. Colliot, T. Mansi, N. Bernasconi, V. Naessens, D. Klironomos, and a Bernasconi, “Segmentation of focal cortical dysplasia lesions on MRI using level set evolution.,” *Neuroimage*, vol. 32, no. 4, pp. 1621–30, Oct. 2006.

[49] D. Bathula and X. Papademetris, “LEVEL SET BASED CLUSTERING FOR ANALYSIS OF FUNCTIONAL MRI DATA,” *Biomed. Imaging From Nano to Macro, 2007.*, vol. 4, no. 4193311, pp. 416–419, 2007.

[50] W. Yu, H. Lee, S. Hariharan, and W. Bu, “Level set segmentation of cellular images based on topological dependence,” *Adv. Vis.*, pp. 540–551, 2008.

[51] H. Chang and B. Parvin, “Multiphase level set for automated delineation of membrane-bound macromolecules,” in *Biomedical Imaging: From Nano to Macro, 2010*, 2010, pp. 165–168.

[52] A. Sofou and P. Maragos, “Generalized flooding and Multicue PDE-based image segmentation,” *IEEE Trans. Image Process.*, vol. 17, no. 3, pp. 364–376, 2008.

[53] A. Sofou and P. Maragos, “PDE-based modeling of image segmentation using volumic flooding,” *Image Process. 2003. ICIP 2003.*, no. September, pp. 431–434, 2003.

[54] JEOL, “A Guide to Scanning Microscope Observation.”

[55] R. Redondo, G. Bueno, G. Cristóbal, J. Vidal, O. Déniz, M. García-Rojo, C. Murillo, F. Relea, and J. González, “Quality evaluation of microscopy and scanned histological images for diagnostic purposes.,” *Micron*, vol. 43, no. 2–3, pp. 334–43, Feb. 2012.

[56] P. M. Nederlof, S. van der Flier, a K. Raap, and H. J. Tanke, “Quantification of inter- and intra-nuclear variation of fluorescence in situ hybridization signals.,” *Cytometry*, vol. 13, no. 8, pp. 831–8, Jan. 1992.

[57] J. C. Waters, “Accuracy and precision in quantitative fluorescence microscopy.,” *J. Cell Biol.*, vol. 185, no. 7, pp. 1135–48, Jun. 2009.

[58] NIKON, “Nikon Perfect Focus System,” 2012. .

[59] J. Bernsen, “Dynamic thresholding of gray level images,” in *International Conference on Pattern Recognition*, 1986, pp. 1251–1255.

[60] N. Otsu, “A threshold selection method from gray-level histogram,” *IEEE Trans. Syst. Man Cybern.*, vol. 9, pp. 62–66, 1979.

[61] R. Medina-Carnicer, F. Madrid-Cuevas, A. Carmona-Poyato, and R. Muñoz-Salinas, “On candidates selection for hysteresis thresholds in edge detection,” *Pattern Recognit.*, vol. 42, no. 7, pp. 1284–1296, 2008.

[62] K. Yan and J. F. Verbeek, “Segmentation for High-throughput Image Analysis: Watershed Masked Clustering,” in *ISoLA 2012, Part II, LNCS*, 2012, vol. 7610, pp. 25–41.

[63] “Fiji.” [Online]. Available: http://fiji.sc/wiki/index.php/Fiji.

[64] J. Schindelin, I. Arganda-Carreras, E. Frise, V. Kaynig, M. Longair, T. Pietzsch, S. Preibisch, C. Rueden, S. Saalfeld, B. Schmid, J.-Y. Tinevez, D. J. White, V. Hartenstein, K. Eliceiri, P. Tomancak, and A. Cardona, “Fiji: an open-source platform for biological-image analysis,” *Nat. Methods*, vol. 9, no. 7, pp. 676–682, Jun. 2012.

[65] T. J. Collins, “ImageJ for microscopy,” *Biotechniques*, vol. 43, pp. 25–30, 2007.

[66] M. Sezgin and B. Sankur, “Survey over image thresholding techniques and quantitative performance evaluation,” *J. Electron. Imaging*, vol. 13, no. 1, p. 146, 2004.

[67] S. J. Osher and R. P. Fedkiw, *Level Set Methods and Dynamic Implicit Surfaces*. Springer-Verlag, 2002.

[68] R. Malladi, J. A. Sethian, and B. C. Vemuri, “A Topology Independent Shape Modeling Scheme,” in *SPIE Conf. on Geometric Methods in Computer Vision II*, 1993.

[69] J. Canny, “A computational approach to edge detection.,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 8, no. 6, pp. 679–98, Jun. 1986.

[70] Y. Qin, G. Stokman, K. Yan, S. Ramaiahgari, F. Verbeek, M. de Graauw, B. van de Water, and L. Price, “Cyclic AMP signalling protects proximal tubule epithelial cells from cisplatin-induced apoptosis via activation of Epac,” *Br. J. Pharmacol.*, 2011.

[71] J. B. Roerdink and A. Meijster, “The Watershed Transform: Definitions, Algorithms and Parallelization Strategies,” *Fundam. Inform.*, pp. 187–228, 2000.

[72] W. Cai, S. Chen, and D. Zhang, “Fast and robust fuzzy c-means clustering algorithms incorporating local information for image segmentation,” *Pattern Recognit.*, vol. 40, no. 3, pp. 825–838, 2007.

[73] J. Fan, M. Han, and J. Wang, “Single point iterative weighted fuzzy C-means clustering algorithm for remote sensing image segmentation,” *Pattern Recognit.*, vol. 42, no. 11, pp. 2527–2540, 2009.

[74] L. Ma and R. Staunton, “A modified fuzzy C-means image segmentation algorithm for use with uneven illumination patterns,” *Pattern Recognit.*, vol. 40, no. 11, pp. 3005–3011, 2007.

[75] J. Angulo and B. Schaack, “Morphological-based adaptive segmentation and quantification of cell assays in high content screening,” in *Biomedical Imaging: From Nano to Macro, 2008. ISBI 2008. 5th IEEE International Symposium on*, 2008, no. c, pp. 360–363.

[76] F. J. Verbeek, “Three dimensional reconstruction from serial sections including deformation correction,” Delft, The Netherlands, 1995.

[77] F. J. Verbeek, “Theory & Practice of 3D-reconstructions from serial sections,” in in *Image Processing, A Practical Approach*, Oxford University Press, 1999.

[78] P. vander Putten, L. F. M. Bertens, J. Liu, F. Hagen, T. Boekhout, and F. J. Verbeek, “Classification of Yeast Cells from Image Features to Evaluate Pathogen Conditions,” in *SPIE 6506*, 2007.

[79] K. Yan, L. F. M. Bertens, and F. J. Verbeek, “Image Registraton and Realignment using Evolutionary Algorithms with High resolution 3D model from Human Liver,” in *CGIM 2010*, 2010.

[80] C. Xue, J. Wyckoff, F. Liang, M. Sidani, S. Violini, K.-L. Tsai, Z.-Y. Zhang, E. Sahai, J. Condeelis, and J. E. Segall, “Epidermal growth factor receptor overexpression results in increased tumor cell motility in vivo coordinately with enhanced intravasation and metastasis.,” *Cancer Res.*, vol. 66, no. 1, pp. 192–7, Jan. 2006.

[81] T. Bushberg, Jerrold, J. A. Seibert, M. Leidholdt, Edwin, and M. Boone, John, *The Essential Physics of Medicial Imaging*, 2nd ed. Lippincott Williams & Wilkins, 2001.

[82] A. Carpenter, T. Jones, M. Lamprecht, C. Clarke, I. Kang, O. Friman, and E. Al., “CellProfiler: image analysis software for identifying and quantifying cell phenotypes,” *Genome Biol.*, vol. 7, no. 10, 2006.

[83] J. Pu, C. D. McCaig, L. Cao, Z. Zhao, J. E. Segall, and M. Zhao, “EGF receptor signalling is essential for electric-field-directed migration of breast cancer cells.,” *J. Cell Sci.*, vol. 120, no. Pt 19, pp. 3395–403, Oct. 2007.

[84] J. Moffat, D. a Grueneberg, X. Yang, S. Y. Kim, A. M. Kloepfer, G. Hinkle, B. Piqani, T. M. Eisenhaure, B. Luo, J. K. Grenier, A. E. Carpenter, S. Y. Foo, S. a Stewart, B. R. Stockwell, N. Hacohen, W. C. Hahn, E. S. Lander, D. M. Sabatini, and D. E. Root, “A lentiviral RNAi library for human and mouse genes applied to an arrayed viral high-content screen.,” *Cell*, vol. 124, no. 6, pp. 1283–98, Mar. 2006.

[85] A. R. Webb, *Statistical Pattern Recognition*, 2nd ed. Wiley, 2005.

[86] C. J. van Rijsbergen, *Information Retrieval*. 1979.

[87] D. G. Altman and J. M. Bland, “Statistics Notes: Diagnostic tests 1: sensitivity and specificity,” *BMJ*, vol. 308, no. 1552, 1994.

[88] B. J. Pawley, “Critical Aspects of Fluorescence Confocal Microscopy,” *Nikon Microscopy U*, 2012. [Online]. Available: http://www.microscopyu.com/articles/confocal/pawley39steps.html.

[89] R. B. Paranjape, W. M. Morrow, and R. B. Rangayyan, “Adaptive Neighbourhood Histogram Equalization for Image Enhancement,” *CVGIPGMIP*, vol. 54, no. 3, pp. 259–267, 1992.

[90] G. Benedetti, M. Fokkelman, K. Yan, L. Fredriksson, B. Herpers, J. Meerman, B. van de Water, and M. de Graauw, “The NF-κB family member RelB Facilitates Apoptosis of Renal Epithelial Cells Caused by Cisplatin/TNFα Synergy by Suppressing an EMT-like Phenotypic Switch,” *Mol. Pharmacol.*, 2013.

[91] Y. L. Wang, “Exchange of actin subunits at the leading edge of living fibroblasts: possible role of treadmilling.,” *J. Cell Biol.*, vol. 101, no. 2, pp. 597–602, Aug. 1985.

[92] M. Bretscher, “Distribution of receptors for transferrin and low density lipoprotein on the surface of giant HeLa cells,” *Proc. Natl. Acad.*, vol. 80, no. January, pp. 454–458, 1983.

[93] M. Lee, “Human body tracking with auxiliary measurements,” *Anal. Model. Faces Gestures*, 2003.

[94] Y. Huang and S. Huang, Thomas, “2-D Model-based human body tracking,” *Object Recognit. Support. by user Interact. Serv. Robot.*, vol. 1, pp. 552–555.

[95] J. Wang, H. Man, and Y. Yin, “Tracking human body by using particle filter Gaussian process Markov-switching model,” *2008 19th Int. Conf. Pattern Recognit.*, pp. 1–4, Dec. 2008.

[96] B. Gorry, Z. Chen, K. Hammond, A. Wallace, and G. Michaelson, “Using Mean-Shift Tracking Algorithms for Real-Time Tracking of Moving Images on an Autonomous Vehicle Testbed Platform,” *Proc. World*, vol. 25, no. November, pp. 356–361, 2007.

[97] K. Althoff, J. Degerman, C. Wahlby, T. Thorlin, J. Faijerson, P. S. Eriksson, and T. Gustavsson, “Time-Lapse Microscopy and Classification of in Vitro Cell Migration Using Hidden Markov Modeling,” *2006 IEEE Int. Conf. Acoust. Speed Signal Process. Proc.*, vol. 5, pp. V–1165–V–1168, 2006.

[98] K. Jaqaman, D. Loerke, M. Mettlen, H. Kuwata, S. Grinstein, S. L. Schmid, and G. Danuser, “Robust single-particle tracking in live-cell time-lapse sequences,” *Nat. Methods*, vol. 5, pp. 695–702, 2008.

[99] Y. Cheng, “Mean shift, mode seeking, and clustering,” *Pattern Anal. Mach. Intell. IEEE*, vol. 17, no. 8, 1995.

[100] A. Francois, “Real-time multi-resolution blob tracking,” pp. 1–10, 2004.

[101] K. Li, E. D. Miller, M. Chen, T. Kanade, L. E. Weiss, and P. G. Campbell, “Cell population tracking and lineage construction with spatiotemporal context.,” *Med. Image Anal.*, vol. 12, no. 5, pp. 546–66, Oct. 2008.

[102] J. Huth, M. Buchholz, J. M. Kraus, M. Schmucker, G. von Wichert, D. Krndija, T. Seufferlein, T. M. Gress, and H. a Kestler, “Significantly improved precision of cell migration analysis in time-lapse video microscopy through use of a fully automated tracking system.,” *BMC Cell Biol.*, vol. 11, p. 24, Jan. 2010.

[103] A. Hand, T. Sun, D. Barber, and D. Hose, “Automated tracking of migrating cells in phase-contrast video microscopy sequences using image registration,” *J. Microsc.*, vol. 234, no. November 2008, pp. 62–79, 2009.

[104] D. Klopfenstein and R. Vale, “The lipid binding pleckstrin homology domain in UNC-104 kinesin is necessary for synaptic vesicle transport in Caenorhabditis elegans,” *Mol. Biol. Cell*, vol. 15, no. August, pp. 3729–3739, 2004.

[105] C. Yang and R. Duraiswami, “Mean-shift analysis using quasiNewton methods,” *Process. 2003. ICIP*, no. 1, pp. 1–4, 2003.

[106] D. Comaniciu, “Mean shift: A robust approach toward feature space analysis,” *Anal. Mach. Intell.*, vol. 24, no. 5, pp. 603–619, 2002.

[107] B. Gorry, Z. Chen, K. Hammond, and A. Wallace, “Using Mean-Shift Tracking Algorithms for Real-Time Tracking of Moving Images on an Autonomous Vehicle Testbed Platform,” *Proc. World*, 2007.

[108] Y. SUN, C. LIN, C. KUO, and C. HO, “Live cell tracking based on cellular state recognition from microscopic images,” *J. Microsc.*, vol. 235, no. April 2008, pp. 94–105, 2009.

[109] I. F. Sbalzarini and P. Koumoutsakos, “Feature point tracking and trajectory analysis for video imaging in cell biology,” *J. Struct. Biol.*, vol. 151, no. 2, pp. 182–195, 2005.

[110] Z. Khan, T. Balch, and F. Dellaert, “Efficient particle filter-based tracking of multiple interacting targets using an mrf-based motion model,” *Proc. 2003 IEEE/RSJ Int. Conf. Intell. Robot. Syst. (IROS 2003) (Cat. No.03CH37453)*, vol. 1, pp. 254–259, 2003.

[111] E. A. Cavalcanti-Adam, T. Volberg, A. Micoulet, H. Kessler, B. Geiger, and J. P. Spatz, “Cell spreading and focal adhesion dynamics are regulated by spacing of integrin ligands,” *Biophys. J.*, vol. 92, no. 8, pp. 2964–2974, 2007.

[112] C. Chen, “Cell shape provides global control of focal adhesion assembly,” *Biochem. Biophys. Res. Commun.*, vol. 307, no. 2, pp. 355–361, Jul. 2003.

[113] F. Yin and D. Makris, “Performance evaluation of object tracking algorithms,” in *International Workshop on Performance Evaluationon 2007*, 2007.

[114] M. Sidani, D. Wessels, G. Mouneimne, M. Ghosh, S. Goswami, C. Sarmiento, W. Wang, S. Kuhl, M. El-Sibai, J. M. Backer, R. Eddy, D. Soll, and J. Condeelis, "Cofilin determines the migration behavior and turning frequency of metastatic cancer cells," *J. Cell Biol.*, vol. 179, no. 4, pp. 777–791, 2007.

[115] A. Neri and G. L. Nicolson, "Phenotypic drift of metastatic and cell-surface properties of mammary adenocarcinoma cell clones during growth in vitro.," *Int. J. cancer J. Int. du cancer*, vol. 28, no. 6, pp. 731–738, 1981.

[116] M. Huigsloot, I. B. Tijdens, G. J. Mulder, and B. van de Water, "Differential regulation of doxorubicin-induced mitochondrial dysfunction and apoptosis by Bcl-2 in mammary adenocarcinoma (MTLn3) cells.," *J. Biol. Chem.*, vol. 277, no. 39, pp. 35869–79, Sep. 2002.

[117] M. D. Bootman, K. Rietdorf, T. Collins, S. Walker, and M. Sanderson, "Ca2+-sensitive fluorescent dyes and intracellular Ca2+ imaging.," *Cold Spring Harb. Protoc.*, vol. 2013, no. 2, pp. 83–99, Feb. 2013.

[118] T. Schroeder, "Tracking hematopoiesis at the single cell level.," *Ann. N. Y. Acad. Sci.*, vol. 1044, pp. 201–9, Jun. 2005.

[119] H. Samet and M. Tamminen, "Efficient Component Labeling of Images of Arbitrary Dimension Represented by Linear Bintrees," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 10, no. 4, pp. 579–586, 1988.

[120] L. Shapiro and G. Stockman, "Binary Image Analysis," in in *Computer Vision*, Prentice Hall, 2002, pp. 63–105.

[121] N. Harder, F. Mora-Bermúdez, W. J. Godinez, A. Wünsche, R. Eils, J. Ellenberg, and K. Rohr, "Automatic analysis of dividing cells in live cell movies to detect mitotic delays and correlate phenotypes in time.," *Genome Res.*, vol. 19, no. 11, pp. 2113–24, Nov. 2009.

[122] P. Kankaanpää, L. Paavolainen, S. Tiitta, M. Karjalainen, J. Päivärinne, J. Nieminen, V. Marjomäki, J. Heino, and D. J. White, "BioImageXD: an open, general-purpose and high-throughput image-processing platform," *Nat. Methods*, vol. 9, no. 7, pp. 683–689, Jun. 2012.

[123] E. Sahai, "Illuminating the metastatic process.," *Nat. Rev. Cancer*, vol. 7, no. 10, pp. 737–749, 2007.

[124] B. Rosner, *Fundamentals of biostatistics*, 7ed ed. Brooks/Cole: Cengage Learning, 2010.

[125] J. Neyman and E. S. Pearson, "On the problem of the most efficient tests of statistical hypotheses," *Philos. Trans. R. Soc. London Ser. A Contain. Pap. a Math. or Phys. Character*, vol. 231, no. 694–706, pp. 289–337, 1933.

[126] H. Motulsky, *Intuitive Biostatistics*, vol. 17, no. 23. Oxford University Press, 1995, pp. 2804–2805.

[127] G. Shorack and J. Wellner, *Empirical Processes with Applications to Statistics*. 1986, p. p. 239.

[128] A. V. Nefian, L. Liang, X. Pi, X. Liu, and K. Murphy, "Dynamic Bayesian Networks for Audio-Visual Speech Recognition," *EURASIP J. Adv. Signal Process.*, vol. 2002, no. 11, pp. 1274–1288, 2002.

[129] G. Zweig and S. Russell, "Speech recognition with dynamic Bayesian networks," 1998.

[130] L. Xie and H. Yang, "Dynamic Bayesian Network Inversion for Robust Speech," no. 7, pp. 1117–1120, 2007.

[131] B. Alberts, A. Johnson, J. Lewis, M. Raff, K. Roberts, and P. Walter, *Molecular Biology of the Cell*, 4ed ed. New York: Garland Science, 2002.

[132] a a Cohen, N. Geva-Zatorsky, E. Eden, M. Frenkel-Morgenstern, I. Issaeva, a Sigal, R. Milo, C. Cohen-Saidon, Y. Liron, Z. Kam, L. Cohen, T. Danon, N. Perzov, and U. Alon, "Dynamic proteomics of individual cancer cells in response to a drug.," *Science*, vol. 322, no. 5907, pp. 1511–6, Dec. 2008.

[133] S. A. Latt, G. Stetten, L. A. Juergens, H. F. Willard, and C. D. Scher, "Recent developments in the detection of deoxyribonucleic acid synthesis by 33258 Hoechst fluorescence.," *J. Histochem. Cytochem. Off. J. Histochem. Soc.*, vol. 23, no. 7, pp. 493–505, 1975.

[134] M. E. Berginski, E. a Vitriol, K. M. Hahn, and S. M. Gomez, “High-resolution quantification of focal adhesion spatiotemporal dynamics in living cells.,” *PLoS One*, vol. 6, no. 7, p. e22025, Jan. 2011.

[135] D. C. Worth and M. Parsons, “Advances in imaging cell-matrix adhesions.,” *J. Cell Sci.*, vol. 123, no. Pt 21, pp. 3629–38, Nov. 2010.

[136] “ImagePro.” [Online]. Available: http://www.mediacy.com/index.aspx?page=IPP.

[137] S. Sternberg, “Biomedical Image Processing,” *IEEE Comput.*, 1983.

[138] J. G. Lock, B. Wehrle-Haller, and S. Strömblad, “Cell-matrix adhesion complexes: master control machinery of cell migration.,” *Semin. Cancer Biol.*, vol. 18, no. 1, pp. 65–76, Feb. 2008.

[139] S. Le Dévédec, B. Geverts, H. de Bont, K. Yan, J. F. Verbeek, A. Houtsmuller, and B. van de Water, “The residence time of focal adhesion kinase (FAK) and paxillin at focal adhesions in renal epithelial cells is determined by adhesion size, strength and life cycle status,” *J. Cell Sci.*, 2012.

[140] A. E. Carlsson and D. Sept, “Mathematical modeling of cell migration.,” *Methods Cell Biol.*, vol. 84, no. 07, pp. 911–37, Jan. 2008.

[141] E. L. Barnhart, K. Lee, K. Keren, A. Mogilner, and J. A. Theriot, “An Adhesion-Dependent Switch between Mechanisms That Determine Motile Cell Shape,” *PLoS Biol.*, vol. 9, no. 5, 2011.

[142] J. Lee, W. A. Nicewander, and J. Rodgers, “Thirteen ways to look at the correlation coefficient,” *Am. Stat.*, vol. 42, no. 1, pp. 59–66, 1988.

[143] K. Meier, “Intuitive Biostatistics : Choosing a statistical test paired groups.” pp. 1–5, 2004.

[144] H. Lilliefors, “On the Kolmogorov–Smirnov test for normality with mean and variance unknown,” *J. Am. Stat. Assoc.*, vol. 62, pp. 399–402, 1967.

[145] K. Yan, E. Larios, and F. Verbeek, “Automation in Cytomics: Systematic Solution for Image Analysis and Management in High Throughput Sequences,” in *IEEE Conf. Engineering and Technology (CET 2011)*, 2011, pp. 195–198.

[146] E. Larios, Y. Zhang, K. Yan, Z. Di, S. Le Dévédec, S. Groffen, and F. Verbeek, “Automation in Cytomics: A Modern RDBMS Based Platform for Image Analysis and Management in High-Throughput Screening Experiments,” in *Conf. Health Information Science 2012*, 2012.

[147] ITU, “ITU-T X-Series Recommendations,” 2012. [Online]. Available: http://www.itu.int/rec/T-REC-X/en.

[148] B. Beizer, *Black-Box Testing: Techniques for Functional Testing of Software and Systems*. 1995.

[149] “Chapter 2 The Benefits of Modular Programming,” in in *Netbean Users Guide*, Sun Microsystems, 2007.

[150] “Web Services Architecture,” *W3C*, 2004. [Online]. Available: http://www.w3.org/TR/ws-arch/.

[151] M. Ivanova and M. Kersten, “An architecture for recycling intermediates in a column-store,” *ACM Trans. …*, 2010.

[152] P. Boncz, “Breaking the memory wall in MonetDB,” *Commun. ACM*, no. September, pp. 77–85, 2008.

[153] V. F. Van Der Heijden and D. De Ridder, *Classification, parameter estimation, and state estimation: an engineering ...*, vol. 32, no. 2. John Wiley and Sons, 2004, pp. 194–194.

[154] C. Camacho and T. Madden, “SOAP-based BLAST Web Service,” *BLAST Help*, 2011.

[155] J. Quackenbush, “Microarray data normalization and transformation.,” *Nat. Genet.*, vol. 32 Suppl, no. december, pp. 496–501, 2002.

[156] A. Fujita, J. R. Sato, L. D. O. Rodrigues, C. E. Ferreira, and M. C. Sogayar, “Evaluating different methods of microarray data normalization,” *BMC Bioinformatics*, vol. 7, no. 1, p. 469, 2006.

[157] Y. Zhao, M.-C. Li, and R. Simon, “An adaptive method for cDNA microarray normalization,” *BMC Bioinformatics*, vol. 6, no. 1, p. 28, 2005.

[158] P. Carvalho, T. Oliveira, L. Ciobanu, F. Gaspar, L. F. Teixeira, R. Bastos, J. S. Cardoso, M. S. Dias, and L. Côrte-Real, "Analysis of object description methods in a video object tracking environment," *Mach. Vis. Appl.*, Jun. 2013.

[159] Y. Zhong and B. Ji, "Impact of cell shape on cell migration behavior on elastic substrate.," *Biofabrication*, vol. 5, no. 1, p. 015011, Mar. 2013.

Nederlandse Samenvatting

In het onderzoeksveld van de Cytomics staat het begrijpen van cellulair en sub-cellulair gedrag centraal. Om dergelijk onderzoek op een grote schaal te kunnen uitvoeren is automatisering noodzakelijk. Het doel van een experiment in cytomics is het bestuderen van bepaald gedrag van cellen uit beelden die worden verkregen van een geautomatiseerde microscoop opstelling; bijvoorbeeld voor het bestuderen en meten van de migratie van een kanker cel. Dit specifieke doel wordt gerealiseerd door gebruik te maken van geschakelde beeldanalyse procedures. Het doel van deze beeldanalyse is om numerieke kenmerken uit de reeksen van beelden te halen met behulp van digitale beeldverwerkingstechnieken. Wanneer dit wordt toegepast op grote schaal op beeldreeksen afkomstig uit bio-experimenten dan spreekt men vaak van een “high-content screening” (HCS) experiment. Als we het beeldanalyse probleem van een HCS-experiment vergelijken met andere vraagstukken uit de beeldanalyse, dan zien we dat beeldreeksen die worden geproduceerd door HCS-experimenten zoals toegepast in studies voor cytomics, objecten bevatten die significant variëren in intensiteit en vorm. Standaard oplossingen voor de beeldanalyse houden meestal geen rekening met inter-object variatie hetgeen foute resultaten introduceert. Met deze vaststellingen in gedachte ligt het focus van ons onderzoek op het ontwerpen van specifieke oplossingen voor de beeldanalyse die kunnen omgaan met de inter-object variatie in HCS-beeldreeksen waarbij een strategie wordt gebruikt in dewelke de berekeningen zich aanpassen aan de variatie in de data. In dit proefschrift worden specifieke oplossingen behandeld voor twee belangrijke procedures in de beeldanalyse van HCS-beeldreeksen, te weten, (1) beeldsegmentatie en (2) object-volgen (beter bekend als object-tracking).

(1) Beeldsegmentatie is de procedure waarbij objecten, zoals bijvoorbeeld cellen, in een beeld worden herkend. Het is vaak lastig om, gebruik makend van een standaard oplossing voor beeldsegmentatie, de instellingen voor deze procedure zo te kiezen dat er rekening wordt gehouden met de variatie tussen de individuele objecten.

In het onderzoek beschreven in dit proefschrift, wordt de ontwikkeling behandeld van een adaptief algoritme, het zogenaamde “watershed-masked clustering” (WMC) algoritme. In vergelijking met andere algoritmes voor segmentatie laten we zien dat het WMC-algoritme een zeer goede en stabiele resultaten geeft in HCS-studies.

(2) Object-volgen (object-tracking) is een procedure waarbij dynamische informatie wordt verzameld en geëxtraheerd van objecten in een beeldreeks; bijvoorbeeld de snelheid van migratie van een cel. Een algoritme voor object tracking construeert links tussen objecten uit opeenvolgende beelden uit een beeldreeks waarbij gebruikt wordt gemaakt van informatie over de object-vorm en/of de object-positie. In het onderzoek in de celbiologie zijn de objecten cellen of sub-cellulaire structuren. Objecten per beeld in (ongeveer) dezelfde positie zijn en/of vergelijkbare vorm hebben worden aan elkaar gerelateerd. Omdat het een tijdreeks betreft kan er zo een “traject” worden vastgesteld en uit dit traject kunnen metingen zoals snelheid en bewegingspatroon worden berekend.

In het onderzoek beschreven in dit proefschrift worden twee “tracking” algoritmes geïntroduceerd, namelijk het algoritme bekend als “*kernel density estimation with mean shift*” en het algoritme bekend als “*energy driven linear motion*”. Onze testen hebben uitgewezen dat deze twee algoritmen accurate “tracking” resultaten genereren. Dit wordt eveneens bevestigd door de twee case studies die in dit proefschrift zijn gepresenteerd.

De volgende twee case studies laten de bruikbaarheid van onze specifieke oplossingen voor beeldanalyse zien.

Case-studie 1: Onze beeldanalyse oplossing heeft het mogelijk gemaakt, metingen te verkrijgen voor motiliteit en morfologie van kanker cellen die behandeld waren met een groei-factor. De metingen kunnen worden gebruikt om de verschillen te onderscheiden in het gedrag van de cel zoals dat wordt geïnduceerd door verschillende groei-factoren. Met de metingen kan een bijdrage worden geleverd aan het doorgronden van respons die door een medicijn wordt geïnduceerd uitgedrukt in motiliteit en morfologie.

Case-studie 2: In deze case-studie laat onze beeldanalyse oplossing de mogelijkheid zien om voor een sub-cellulair complex metingen te extraheren uit een beeldreeks. In het bijzonder wordt hiermee de dynamiek van matrix adhesie eiwitten die betrokken zijn bij de beweging van de cel geïllustreerd. De analyse uit deze case-studie bevestigt op numerieke wijze dat de omzet van deze matrix-adhesie complexen, uitgedrukt in duur van assemblage en disassemblage, sterk geassocieerd is met de regulatie van cel motiliteit.

Met het beschikbaar hebben van een goed systeem voor data-management, kan de beeldanalyse op een efficiënte manier worden uitgevoerd. Het systeem voor datamanagement schermt de eindgebruiker af van de onderliggende complexiteit van de berekeningen in de beeldanalyse door een hoogwaardig grafisch gebruikers-interface (GUI). Het geautomatiseerde proces onderliggend aan het gebruikers interface maakt het mogelijk om te schalen naar een groter volume van beelddata. Voorts neemt, door dit systeem, de toegankelijkheid van de interdisciplinaire data toe doordat een standaardisatie is doorgevoerd van data formaten en opslag van meetgegevens tussen verschillende experimenten. Deze standaardisatie van data formaten is een belangrijk kenmerk omdat een uiteindelijk doel van cytomics is, een integratie tot stand te brengen met andere –omics data zodat de correlatie tussen genotype en fenotype bestudeerd kan worden.

Concluderend, het onderzoek dat in dit proefschrift wordt beschreven heeft tot doel efficiënte en robuuste oplossingen voor HCS analyse te ontwerpen. Door bestaande algoritmen te analyseren en bestuderen, hebben we algoritmen kunnen ontwerpen die goed passen bij de unieke karakteristieken van een HCS experiment. De case-studies laten zien dat de beeldanalyse oplossingen die zijn geïmplementeerd waarbij deze algoritmen worden gebruikt, een goed platform bieden voor het verkrijgen van objectieve informatie voor het doorgronden van biologische vraagstukken. Vergeleken met handmatige analyse kunnen de geautomatiseerde oplossingen voor beeldanalyse de HCS analyse verder objectiveren. Op deze manier wordt de weg bereid voor het begrijpen van de controle mechanismen die het gedrag van de cel bepalen.

English Summary

Cytomics is the study to understand cellular or subcellular behavior. In order to do research in cytomics on a large scale, automation is needed. The goal of a cytomics experiment is to study the cell behavior from images captured with an automated microscope setup; for example measuring migration of a cancer cell. The goal is achieved by using an image analysis pipeline. The aim of image analysis is to extract numerical descriptors from image sequences using digital image processing. When applied on a large-scale to images from bio-experiments, this is often referred as a high-content screening (HCS) experiment. Compared to other image analysis problems, HCS experiments as employed in cytomics studies often produce image sets containing objects significantly varied in both intensity and shape. Generic image analysis solutions often overlook the between-objects variation, thereby producing false results. To that end, our research focuses on designing dedicated image analysis solutions to cope with the between-object variation in HCS images using a strategy through which the computation adapts to the variation in the data. In this thesis, dedicated solutions for two procedures, namely image segmentation and object tracking, in the image analysis of HCS experiment are illustrated:

(1) Image segmentation is the procedure to extract objects from an image, i.e. cells. Often it is difficult to tune the parameters of a generic segmentation solution to the variation between individual images.

In the research described in this thesis, a self-adaptive segmentation algorithm, namely the Watershed Masked Clustering (WMC) algorithm has been developed. Compared to other algorithms, the WMC that we developed has demonstrated a robust performance in HCS studies.

(2) Object tracking is a procedure that will extract dynamic information from the objects, i.e. speed of migration. An object tracking algorithm builds linkages between objects from consecutive frames using information such as object shape and/or position. In cell biology studies, such an object can be a cell or a subcellular structure. Objects that are in proximity over frames and/or of similar shape will be related. From the trajectory, measurements such as a velocity or motion pattern can be extracted.

In the research described in this thesis, two tracking algorithms, namely the kernel density estimation with mean shift algorithm and the energy driven linear modeling algorithm, are introduced. From our testing, these two algorithms produce accurate tracking results in both case studies illustrated in this thesis.

The following two case studies demonstrate the applicability of our dedicated image analysis solutions:

Case Study 1: Our image analysis pipeline was capable of extracting both motility and morphology measurements from growth factor treated cancer cell. These measurements can be used to distinguish the changes in cell behavior induced by different growth factors. These measurements can further contribute to an objective understanding of drug-induced responses in terms of cell motility and morphology.

Case Study 2: In this case study, our image analysis pipeline demonstrated the possibility to extract measurements for subcellular structures in a study on the dynamics of matrix adhesions. The analysis has numerically confirmed that the turnover of matrix adhesions, in terms of assembly and disassembly duration, is strongly associated with the regulation of cell motility.

With a good data management system available, the image analysis can be performed in an efficient manner. The data management system shields end-users from the underlying complexity of the computational approaches in the image analysis by providing an integrated high-end graphic user interface (GUI). The automation underlying the GUI will enable to scale to a higher volume of image data. It further increases the accessibility of interdisciplinary data by standardizing different data formats and measurement structures between experiments. The standardization of data formats is an important feature since the ultimate goal of cytomics is to be integrated with other –oimcs data in order to study the genotype-to-phenotype correlation.

In conclusion, the research described in this thesis aims to design efficient and robust HCS analysis solutions. By studying existing algorithms, several dedicated image analysis algorithms are designed to fit the unique image characteristics of HCS experiment. From the case studies, it shows that analysis pipelines using these dedicated algorithms can well provide a platform of extracting objective understanding of biological questions. Compared to manual analysis, an automated solution will objectivize HCS analysis. It further paves the way in the understanding of control mechanisms of cell behavior.

List of Publications

Benedetti, G., Fokkelman, M., Yan, K., Fredriksson, L., Herpers, B., Meerman, J., Van de Water, B., et al. (2013). The NF-κB family member RelB Facilitates Apoptosis of Renal Epithelial Cells Caused by Cisplatin/TNFα Synergy by Suppressing an EMT-like Phenotypic Switch. *Molecular Pharmacology*.

De Graauw, M., Cao, L., Winkel, L., Martine, M. H. A. M., Le Dévédec, S. E., Klop, M., Yan, K., et al. (2013). Annexin A2 depletion delays EGFR endocytic trafficking via cofilin activation and enhances EGFR signaling and metastasis formation. *Oncogene*. doi: 10.1038/onc.2013.219. [Epub ahead of print]

Larios, E., Zhang, Y., Yan, K., Di, Z., Le Dévédec, S., Groffen, S., & Verbeek, F. (2012). Automation in Cytomics: A Modern RDBMS Based Platform for Image Analysis and Management in High-Throughput Screening Experiments. *Conf. Health Information Science 2012*. Springer.

Le Dévédec, S., Geverts, B., De Bont, H., Yan, K., Verbeek, J. F., Houtsmuller, A., & Van de Water, B. (2012). The residence time of focal adhesion kinase (FAK) and paxillin at focal adhesions in renal epithelial cells is determined by adhesion size, strength and life cycle status. *Journal of cell science*.

Di Z., Herpers B., Fredriksson L., Yan K., van de Water B., et al. (2012) Automated Analysis of NF-κB Nuclear Translocation Kinetics in High-Throughput Screening. *PLoS ONE* 7(12): e52337. doi:10.1371/journal.pone.0052337

Yan, K., & Verbeek, J. F. (2012). Segmentation for High-throughput Image Analysis: Watershed Masked Clustering. *ISoLA 2012, Part II, LNCS* (Vol. 7610, pp. 25–41). Heidelberg: Springer.

Qin, Y., Stokman, G., Yan, K., Ramaiahgari, S., Verbeek, F., De Graauw, M., Van de Water, B., et al. (2011). Cyclic AMP signalling protects proximal tubule epithelial cells from cisplatin-induced apoptosis via activation of Epac. *British Journal of Pharmacology*.

Cao, L., Yan, K., Winkel, L., De Graauw, M., & Verbeek, F. (2011). Pattern Recognition in High-Content Cytomics Screens for Target Discovery - Case Studies in Endocytosis. *Lecture Notes in Computer Science* (pp. 330–342). Springer.

Damiano, L., Le Dévédec, S. E., Di Stefano, P., Repetto, D., Lalai, R., Truong, H., Xiong, J. L., Yan, K., et al. (2011). p140Cap suppresses the invasive properties of highly metastatic MTLn3-EGFR cells via impaired cortactin phosphorylation. *Oncogene*, *30*(5), 624–33. doi:10.1038/onc.2011.257

Yan, K., Larios, E., LeDevedec S. van de Water, B., Verbeek FJ (2011), Automation in Cytomics: Systematic Solution for Image Analysis and Management in High Throughput Sequences. *Proc. IEEE Conf. Engineering and Technology (CET 2011)*, Vol 7. Shanghai 2011, 195-198

Qin, Y., Stokman, G., Yan, K., Ramaiahgari, S., Verbeek, F., van de Water, B., Price, L.S., (2010), "Activation of Epac-Rap Signaling Protects against Cisplatin-induced Apoptosis of Mouse Renal Proximal Tubular Cells, *Drug Metabolism Reviews*, vol. 42, pp. 31-31.

LeDévédec, S., Yan, K., De Bont, H., Ghotra, V., Truong, H., Danen, E., Verbeek, F. J., et al. (2010). A Systems Microscopy Approach to Understand Cancer Cell Migration and Metastasis. *Cellular and Molecular in Life Science*, *67*(19), 3219–3240.

Yan, K., Bertens, L. F. M., & Verbeek, F. J. (2010). Image Registraton and Realignment using Evolutionary Algorithms with High resolution 3D model from Human Liver. In A. D. Sappa (Ed.), *CGIM 2010*. Innsbruck, Austria.

Yan, K., LeDévédec, S., Van De Water, B., & Verbeek, F. (2009). Cell Tracking and Data Analysis of in vitro Tumour Cells from Time-Lapse Image Sequences. *VISAPP 2009* (pp. 281–287). Lisbon.

Curriculum Vitae

Kuan Yan was born on December 30th 1982, in Shanghai, the People's Republic of China. In 2002, he started his bachelor study in computer science at InHolland University of Applied Sciences in Amsterdam. He completed his bachelor degree in June 2006. His bachelor project focused on the platform design for Business-to-Consumer (B2C) online shopping. In 2006, he started a master study in Bioinformatics at Leiden University and completed his Master's degree in May 2008. His master thesis project focused on automated image analysis for the dynamic quantification of MTLn3 rat mammary tumor cell migration from image sequences. The master thesis project was accomplished under the supervision of Dr. Fons J. Verbeek in the section of Imaging and Bioinformatics of LIACS.

In June 2008 Kuan Yan started his PhD in the section Imaging and Bioinformatics of the Leiden Institute of Advanced Computer Science (LIACS) in Leiden University, under the supervision of Dr. Ir. Fons Verbeek; in collaboration with the Toxicology Department of Leiden Academic Center of Drug Research (LACDR); in a project within the BioRange program of the Netherlands Bioinformatics Centre (NBIC). The PhD research focused on the design of dedicated image and data analysis solutions for *in vitro* cancer phenotyping for dynamic HT/HC cell screening.

In 2012, he has been involved in other analysis pipelines for Toxicology Department of LACDR including actin fibers, endoplasmic reticula, mitochondria and nuclear membranes.

As of June 2013, he is employed by OCellO B.V. in the integrated bioinformatics analysis of the HT/HC screens.

Acknowledgements

The journey of my thesis research started in May 2008 and it was ended June 2013. It would not possible to write this thesis without the help and support of people around me. I would like to explicitly mention a few persons in the completion of this journey.

First of all, I would like to thank to my parents who have unconditionally shared their life experiences and academic knowledge with me: You have taught me to be patient and self-controlled regardless the circumstances. In the end, all has turned for the good.

I would like to thank to Ying Shi, my wife, and my parents-in-law who believe in me from the beginning. Thanks for your dedicated support.

Many thanks to my research group collegeus in the imaging and bioinformatics section: Alexander Nezhinsky, Amalia Kallergi, Dome Potikanond, Enrique Larios, Irene Martorelli, Joris Slob, Laura Bertens, Lu Cao, Mohammed Tlais, Rafael Carvalho, Yang Peng and Yanju Zhang. I will always remember those lunch breaks, BBQs, birthday cakes, Sinterklaas parties and etc. My time at LIACS has been pleasurable amidst you people. Laura, I would like to thank specifically for the Dutch lessons: “Dank je wel!”.

I would like to thank my colleagues at Toxicology Department of LACDR: Hans de Bont, Michiel Fokkelman, Sandra Zovko, Sylvia Le Dévédec, Yu Qin, and Zi Di; and especially Sylvia for sharing her professional knowledge as well as her personal support as a good friend.

I would also like to thank my important friends outside the academic world: Ling, family and friend, I will never forget all the laughter we have shared when you still lived in the Netherlands. It is for certain that you are doing great in Shanghai now. Thank you for all those good moments.

Furthermore, I would like to thank all my former teachers of the master study in bioinformatics at Leiden University. Your dedication to your works made it possible for me to graduate three months earlier.

Last but not least, I would like to thank all my former teachers of the bachelor study in computer science at InHolland University. The results would not have been possible without the foundation you laid down ten years ago.