Quantitative analysis of TMA images using computer vision and machine learning approaches



Haiyue Yu

Supervised by Professor Julia Alison Noble
Lincoln College

University of Oxford

A thesis submitted for the degree of

Doctor of Philosophy

Hilary 2016

This thesis is submitted to the Department of Engineering Science, University of Oxford, in partial fulfilment of the requirements for the degree of Doctor of Philosophy. The thesis is entirely my own work and except where otherwise stated, describes my own research.

Haiyue Yu, Lincoln College

Acknowledgements

I am very grateful to take this opportunity to acknowledge and thank all those who have made it possible for me to study and contribute to this field.

I would like to thank my supervisors: Professor Alison Noble, who guided me throughout my research, and my unofficial supervisor Professor Bass Hassan, who was instrumental in setting the direction of my thesis. My thesis would have been impossible without the help and support of both of my supervisors.

I would like to thank my friends and colleagues Dr Simon Li and Dr Claudia Buehnemann for sharing insights and collecting images. I would also like to thank my friends in the lab, Christopher, Weidi, Ana, Omar and Ruobing for their stimulating conversation and willingness to discuss their work.

Finally, all my love and thanks to my family. I wish to thank my parents for encouraging me in academic work and supporting me throughout my studies. Many friends who have given me encouragement for me, they are too numerous to mention all of them. I'd particularly like to mention Campbell, Ying Bai, Rui Jia, Jin Li, Chao Yan for their support in the final year of this thesis.

Quantitative Analysis of TMA images using computer vision and machine learning approaches

Haiyue Yu, Lincoln College, Oxford
Submitted for the degree of Doctor of Philosophy, Engineering Science
Hilary 2016

Abstract

In the research of cancer biomarker and cell signaling pathways, multiple immunofluorescence labelled tissue microarrays (TMAs) are often examined by pathologists. By looking at TMAs under microscopy, the pathologists can score the images visually. In order to improve the accuracy and reliability of the image scoring, an automatic quantification of TMA image is in high demand.

This thesis presents the development and the validation of several key components of image analysis in order to automatically quantify TMA images. The first component is pre-processing of the image, the uneven illumination in fluorescent confocal microscopic images are corrected using a retrospective method and it is validated on this dataset.

The second component is the design and validation of an original cell segmentation algorithm. This algorithm automatically segments cell nuclei and cell cytoplasm from the images. The validation of this algorithm shows more than 90% of accuracy subject to tissue images with good qualities. The quality of tissue images refers to the problems such as, clumpy and overlaping nuclei, DNA degradation, stromal cells and etc. As the tissue quality varies, the texture of tissue changes. The texture of tissue with different quality was analyzed and a classification model is built to classify the textures.

The third component is an automatic tissue quality assessment algorithm for TMA images. In addition, an original framework is designed to assess the quality of TMA by automatically recognize different tissue types and produce the proportion of each type. This proportion of tissue types provides a quality metric for an image and it is shown in the study that this metric is useful for classify the tissue image quality.

In conclusion, this thesis presents an original the pipeline including the three components, to analyze tissue microarray images. The validation of each method has shown good performance.

Contents

1 Introduction **1**

2 Literature Review **4**

2.1 Introduction 4

2.2 Background 5

2.3 Imaging System 8

2.4 Image Analysis 11

2.4.1 Image Pre-processing 13

2.4.2 Image Segmentation 13

2.4.3 Quantitative Analysis 17

2.4.4 Related Work in Machine Learning and Computer Vision . . . 19

2.5 Tissue Image Analysis 21

2.6 Conclusion 21

3 Pre-processing of Tissue Microarray Images **23**

3.1 Introduction 23

3.2 Explanation of imaging data 24

3.2.1 Accuracy and precision in quantitative analysis 26

3.3 Retrospective Illumination Correction of Confocal Images 30

3.3.1 Literature Review 30

3.3.2 Methodology 32

3.3.3 Image Intensity Saturation Correction 39

3.3.4 Validation 41
3.4 Illumination Correction and AQUA Threshold 43
3.5 Summary 43

4 Image Segmentation of Tissue Micro-array Images 44
4.1 Introduction 44
4.2 Collaborations 45
4.3 Literature Review 45
4.4 Overview of the Segmentation Algorithm 47
4.5 Segmentation Algorithm 49
4.5.1 Segmentation of Nuclei 49
4.5.2 Level Set 49
4.5.3 Cell Detection 53
4.5.4 Segmentation of the cytoplasm 58
4.5.5 Discussion about cell cytoplasm segmentation 60
4.6 Quantitative feature extraction 63
4.7 Validation of the Algorithm 63
4.8 Conclusion 65

5 Classification of patterns in tissue samples with different quality 66
5.1 Introduction 66
5.2 The problem 67
5.3 Methodology 74
5.3.1 Data Annotation 74
5.3.2 Image Features 75
5.3.3 Classification models 78
5.4 Experiment 79
5.4.1 Feature Selection 80

5.4.2 Insight into Image Features 84
5.4.3 Random Forest Classification 89
5.4.4 Validation 91
5.5 Discussion 92
5.6 Conclusion 93

6 Automatic assessment of tissue micro-array quality 94
6.1 Introduction 94
6.2 Literature Review 95
6.3 Methodology 97
6.3.1 Object Proposal and Image Partition 98
6.3.2 Object Classification 100
6.4 Experiments 102
6.4.1 Object classification 102
6.4.2 Quantification of Tissue with Different Quality 107
6.4.3 Image Quality Control 107
6.4.4 Correlation with Nuclei Segmentation 109
6.5 Validation of Tissue Quantification 112
6.6 Discussion 114
6.7 Summary 114

7 Conclusions and Future Work 115
7.1 Future Work 116
7.1.1 Translational Work 116
7.1.2 Deep Learning 116

A 118

Bibliography 121

List of Figures

1.1 The pipeline of TMA image analysis. 2

2.1 The six hallmarks of cancer cells [65]. (1) self-sufficiency in growth signals, (2) insensitive to growth inhibitory signals, (3) evasion of programmed cell death, (4) limitless replication potentials, (5) sustained angiogenesis, (6) tissue invasion and metastasis. 5

2.2 Signalling pathways in cells [65]. Colour indicates different cell processes. 6

2.3 Principal light pathways in confocal microscopy [40]. 9

2.4 Information flow pathway in confocal microscopy [40]. 9

3.1 The flow chart showing the procedures of acquire confocal images. . . 26

3.2 Example of TMAs Images. (a) shows the tissue nuclei (DAPI stained) image is brighter at right side of the image than it is on the left side. (b) shows another example where the left side of the image is brighter than the right side. (c) shows the darker corners on the image which is known as vignetting. This error occurs when the camera shutter speed is too fast to allow the film to respond or it might be a result of field curvature of the camera lens. 28

3.3 An example of the digital image obtained from the microscope using Olympus Fluoview software. (a) the DAPI channel is shown in hi-lo model, where the blue pixels indicate their value being 0, while the red pixels indicate their value being 4095 and the values in between them are showing in a gray scale. (b) an example showing the effect on the image, when the low threshold were changed by biologists. (c) the cell nucleus A (marked by yellow circle) is shown an over-saturated nucleus while another nucleus B (marked in magenta circle) is lower in intensity and showing a pattern of chromosome. The nucleus marked by green circle is a different type of cell showing different pattern. The reason that nucleus A is much brighter than nucleus B might related to their cell phrase. (d) examples of saturation and lost of nuclei pattern because of it. 29

3.4 An example of the retrospective illumination of confocal microscopic images. (a) the observed image acquired by the confocal microscope. (b) An overview of the correction algorithm 34

3.5 An example of the bright and dark region computed by the algorithm. (a) The bright region is labelled using red color and the dark region is labelled with blue color. (b) The quantile of the intensity values of pixels in bright region and dark region. 36

3.6 An example of the optimisation of the Gaussian filter size and σ. (a) An example of empirical quantiles of dark and bright regions after filtered with a 100×100 pixel Gaussian filter. (b) An example of empirical quantiles of dark and bright regions when applied a 500×500 pixel Gaussian filter. (c) An example of empirical quantiles of dark and bright regions when applied a 1900×1900 pixel Gaussian filter. (d) The value of the cost function against filter size (σ is set to half of the filter size). (e) The value of the cost function against σ (filter size is 900 pixels). (f) The quantiles of dark and bright pixels when filtered by the optimised filter size (900 pixels) and σ (150 pixels) 38

3.7 An example of the uneven illumination obtained from the method and the results after the image is corrected by this uneven illumination function. 38

3.8 An example of the saturation correction. (a) The histogram of the original image. (b) The histogram of image after filtering, the number of saturated pixels has decreased comparing with (a). (c) The histogram of pixels after saturation correction is plot against the original intensity distribution. 39

3.9 An example showing the image before and after correction. (a) the original image. (b) the image after correction. 40

3.10 The visual examination of images. The images at the left hand side are the original images. The images at the right hand side are corrected images. 42

4.1 Compare of cellular images of cell core and cultured cell. (a) An example of tissue image. (b) Cells grow 3 dimensionally can have different orientations. (c) This is an example from a well-know cultured cell dataset [35] 47

4.2 Overview of the image segmentation algorithm 48
4.3 An example showing the local interior and local exterior regions [90]. (a) Local interior. (b) Local Exteriror. 50
4.4 An example showing the initialisation of level set affects the segmentation results. The contour of the yellow regions indicates its initialisation. The blue line is the line indicating outer side of the curve while the magenta line indicate the inner side of the curve. (a) The initialisation include part of two nuclei and background pixels. (b) The initialisation locates at the centre of the nucleus. (c) The initialisation located at the background captured background pixels.(d)An example showing the results of level set initiated with different threshold values. The red line indicates the results from an initial of higher intensity values, and the yellow lines indicates the results from a lower initial values. 52
4.5 An example showing the results of level sets initiated with a set intensities. (a) the original image shown in a parula color range, so that it is easy to see the background in the image, which is difficult in a gray color range.(b) An example showing the results after adding 20 level sets together, the nuclei that are in focus appear in all 20 segmentations. 55
4.6 Split Clumps. (a) The schema of splitting algorithm. (b) The histogram of clumps sizes after several iterations. 56
4.7 An example showing the over segmentation of nuclei. (a) an example of over segmentation of nuclei. (b) change of parameter H can smooth the distance map in watershed. (c) after tuning H, the over-segmentations are removed. 56

4.8 An example showing the the segmentation of clumps of nuclei. (a) The original image in gray color map. (b) The initialisation locates at the centre of the nucleus. (c) The initialisation located at the background captured background pixels.(d) An example showing the results of level set initiated with different threshold values. The red line indicates the results from an initial of higher intensity values, and the yellow lines indicates the results from a lower initial values. 58
4.9 An example to show the CD99 staining of the cell membrane. 59
4.10 An example to show the segmentation of cell. The figure at the left shows the segmentation of cell nuclei. The figure in the middle shows the segmentation of cell nuclei and cytoplasm can be mapped on Ki67 biomarker channel. A quantitative measurement can be made to Ki67 based on segmentation results. The figure at right, shows the segmentation of cytoplasm based on equal distance between nuclei. 60
4.11 An example to show the difference of cytoplasm segmentation among experts.(a) the original tissue images. (b) manual segmentation of (a) by expert 1. (c) manual segmentation of (a) by expert 2. 61

5.1 An example of a poor quality TMA. There are a large amount of stormal cells at the left side of the image and background staining. The right bottom corner and left upper corner have no cells and there are not many tumour cells in this image. 68
5.2 An example of a poor quality TMA. There are stromal cells in different regions and the cells form big clumps. 69
5.3 An example of a poor quality TMA. The cell nuclei has less DAPI staining compared with the cell cytoplasms. It might be a result of degraded nuclei. 70

5.4 An example of a tissue image with good quality. There are many tumour cells and few stromal cells in the image and the boundaries between nuclei are clear. 71

5.5 Examples of image patches taken from original tissue images. (a) An example of 'normal' nuclei. (b) Nuclei are clumpy and overlapping. (c) The DNA in the nuclei sample is degraded. (d) The nuclei in this image showed signs of degradation, but seems less severe compared with (c). The size and shape of nuclei showed heterogeneity with them in (c).(e) Background staining and noise. (f) Stromal tissue in the sample is stained by the fluorescent probes. 73

5.6 The sequential feature selection and feature importance of all the variables for image patch Level 1. (a) By adding the features sequentially, the model improved classification accuracy. (b) Feature importance obtained by a random forest model. The number corresponds to the location of the feature in the feature vector in Table. 5.1. 82

5.7 The sequential feature selection and feature importance of all the variables for image patch Level 2. (a) By adding the features sequentially, the model improved classification accuracy. (b) Feature importance obtained by a random forest model. The number corresponds to the location of the feature in the feature vector in Table. 5.1. 83

5.8 The box plot of discriminative features for level 1 patches. 86

5.9 The box plot of discriminative features for level 2 patches. 87

5.10 The box plot of discriminative features for level 2 patches (continued). 88

5.11 The classification scores for level 1 and level 2 image patches. There are 5 sets of 60 data points (level 1) or 150 data points (level 2) sequentially ordered as shown in the (a) and (b). The different colour represent the probability that the data points belong to a particular class. The red colour represents clumpy nuclei class, the yellow colour represents the stromal tissue class, the blue colour represents the degraded nuclei class, the green colour represents the normal nuclei class and the magenta colour represents the background class. 90
5.12 Classification results on 100 new images patches. 91

6.1 The overview of the framework. 98
6.2 An example of the object proposal of an TMA image. (a)The original DAPI channel of the TMA image. This image shows a large amount of cell nuclei form clusters. (b) The partition of the image based on DAPI channel. (c) The image regions obtained from image partition. (d) An example of a variety of sized image patches. 99
6.3 Examples of image patches of tissue samples. 103
6.4 Sequential selection of discriminative image features. The accuracy of RFs vary with different their levels and y-axis are different. The numbers are corresponding to the location in vector in Table 6.1. . . . 105
6.5 Evaluation of the classification on independent dataset. The result is showing the performance of the classifier when it is used to classify the data. (a) Classification results of Level 0. (b) Classification results of Level 1. (c) Classification results of Level 2. 106
6.6 An examples of quantification tissue quality. (a) The original image for analysis. (b) The proportion of each type of tissue in this image found by the algorithm. 107

6.7 More examples of tissue images. (a) an example image. (b) the proportion of each type of tissue in image (a) found by the algorithm. (c) further example image. (d) The proportion of each tissue type found by the algorithm from image (c). 108
6.8 Test results reported as a confusion matrix. The figure shows that 20 good quality images are classified as good quality while 18 bad quality are classified as bad quality. In addition, 6 images that are good quality, but classified as bad quality. 5 images are bad quality and classified as good quality. 109
6.9 The Bland-Altman test of INR by predicted computer and the INR calculated in experiment. 111
6.10 The proportion of each object that are labelled by computer is compared with that of a pathologist. The pie chart of the left hand side is showing the proportion of each type of object identified by computer algorithm. The right hand side is showing the proportion of each type of object identified by a pathologist. 113

Chapter 1

Introduction

Approximately, one in three people are likely to be affected by cancer during their life time [12]. As a result, further research on cancer is important in order to find better treatments. One of the main type of imaging data in cancer diagnosis and treatment is immuno-fluorescence examination where biopsy tissue is sliced into thin sections and immunolocalised for examination under a fluorescence microscope. However, manual examination of millions of cells is time-consuming and subjective. In order to automate and quantitate the analysis of tissue biopsies, researchers have developed computer algorithms and software to quantify microscopic images, which has developed into the field of digital pathology field.

In this thesis, the state-of-art computer vision and machine learning techniques are employed to design fully automatic quantitative tissue microarray image analysis methods which require minimal user input. As shown in Fig. 1.1, this thesis focuses on implementation of an analysis pipeline consists of image pre-processing, image quality control (Chapter 5 and Chapter 6) and quantification (Chapter 4). In this research, all the imaging data are obtained from Ewing Sarcoma patients. The complications in real biopsy images showed the challenges of successful image recognization. The author addresses these challenges and presents practical solutions to overcome them.

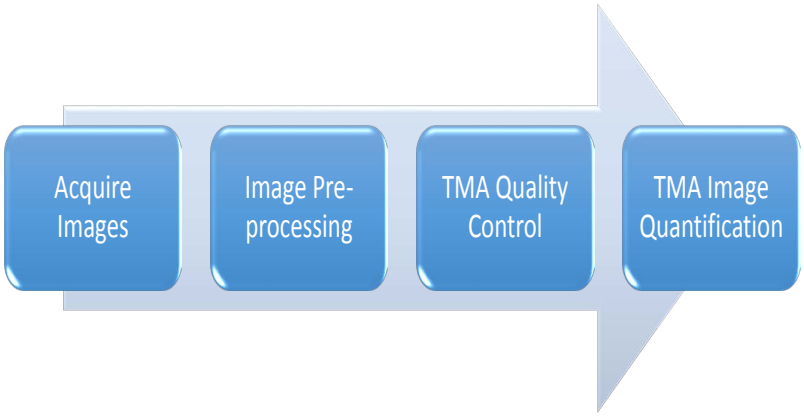


Figure 1.1: The pipeline of TMA image analysis.

In Chapter 2, background literature on cancer biomarker and fluorescent confocal microscopy is reviewed to better understand the fundamental principles of image quantification. Furthermore, previous investigating methods of quantifying microscopic images is discussed in details to highlight the main challenges in digital pathology research today.

In Chapter 3, a practical method is developed to correct the shading effect and systematic errors in confocal microscopic images. This method is validated on our imaging dataset.

In Chapter 4, a fully automatic cell segmentation and quantification method is proposed and validated. Using this method, it is possible to segment individual cells from tissue images and extract features of each cell. This segmentation algorithm is validated against expert-based ground truth and the results show over 90% of accuracy on tissue images with good quality. From this study, it is clear that the tissue microarray quality affects the accuracy of automatic quantification. As a result, this

led to the study of automatic quality control in following chapters.

In Chapter 5, the author is interested in the statistically learning the differences in patterns of different tissue types. Image features such as texture descriptors and intensity etc are investigated in this chapter.

In Chapter 6, a fully automatic Tissue Microarray (TMA) quality control framework is designed. Using this method, the quality of a TMA image can be automatically quantified and a pathologist can select the samples they want to included in their analysis without any bias. Based on this thesis, future work is discussed in the Chapter 7.

Chapter 2

Literature Review

2.1 Introduction

The rapid development of imaging technology is influencing cancer research and generating new tools and solutions. High content screening (HCS) of tissue micro-array is used widely in modern biological laboratories to monitor biological phenomena and it allows researchers to investigate cellular events and to collect quantitative data from complex biological systems. HCS has now been adopted by the pharmaceutical industry [102] and included in all stages of the drug development process. A high throughput high content experiment generates an enormous volume of imaging data, which requires analysis by researchers. The state-of-art for this image analysis generally involves manual process or by eye judgments rather than automated solutions. As the volume of imaging data continues to grow, a manual approach is becoming increasingly impractical, and the introduction of automated methods is essential.

In this review, the state-of-art in high content high throughput technology is reviewed in order to identify the gaps between demands and development. In particular, the application of HCS in cancer research is discussed, which is primarily directing the development of HCS technology.

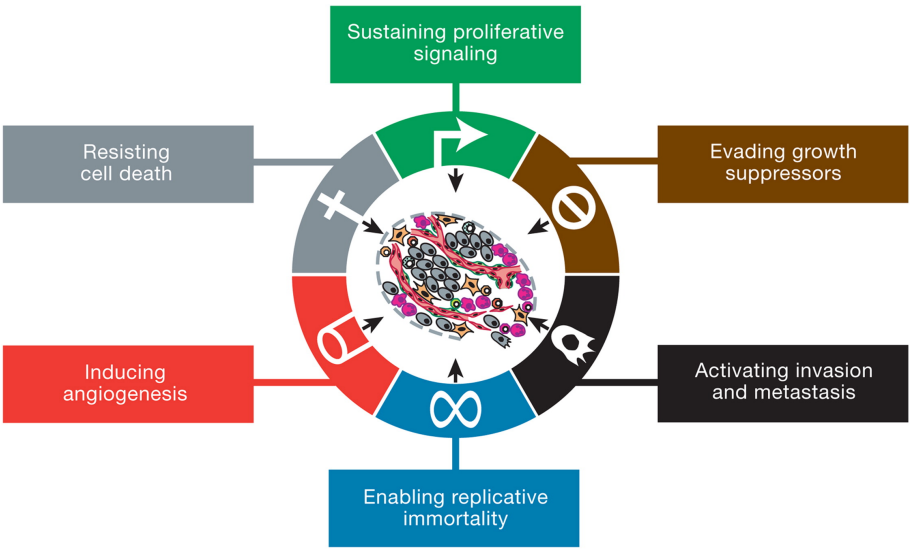


Figure 2.1: The six hallmarks of cancer cells [65]. (1) self-sufficiency in growth signals, (2) insensitive to growth inhibitory signals, (3) evasion of programmed cell death, (4) limitless replication potentials, (5) sustained angiogenesis, (6) tissue invasion and metastasis.

2.2 Background

Hallmark of Cancer Previous research has shown that cancer is a result of malignant growth or the uncontrolled division of cells, which involves dynamic changes in their genome [64]. As shown in the Fig. 2.1, six essential alterations in cells lead to malignant growth.

Cancer research is concerned with understanding: how cancer cells behave in the body, how they interact with the immunology system, their self-regulation abnormalities, and their genome change mechanism [14]. On one hand, the immunology system being harnessed as a tool to destroy cancer cells [136], and on the other molecular biology is trying to identify cancer specific molecules that can be used in drug discovery, cancer detection and prediction. Other research on the genome is trying to understand and repair genes in order to transform cancer cells back into normal (non-cancerous) cells [138]. All of these studies are strongly connected with

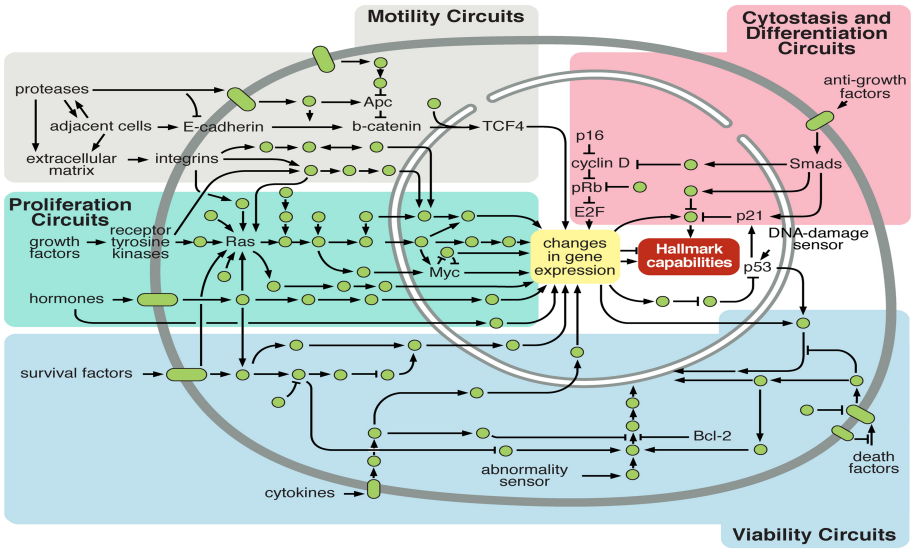


Figure 2.2: Signalling pathways in cells [65]. Colour indicates different cell processes.

signalling pathways, which affect the growth of individual cells and the interactions between cells.

Signalling Pathway in Cells Cell signals help cells to govern their cellular activities and coordinates actions between them. Most cell signals are chemical influence: growth factors, hormones, neurotransmitters, extracellular matrix components, and so on [50]. As shown in Fig. 2.2, the signalling pathways in cells can be represented as an electronic integrated circuit.

The study of signalling pathways supports the discovery of biomarkers [34, 21]. Cancer biomarkers research is a fast growing area, because of its potential for early cancer diagnosis and treatment. There are several types of biomarkers, such as prognostic biomarkers (indicating the progression of disease with or without treatment), pharmacodynamic biomarkers (indicating certain pharmacological responses), diagnostic biomarkers (indicating the diagnose of disease), predictive biomarkers (indicating the most likely responses to certain treatment), and so on.

Cancer biomarker Cancer biomarker translational research is attempting to validate the use of biomarkers for drug discovery applications [47, 153]. However, among several protein and genomic markers, only a few have been advanced to clinical trials. In a given tumour there are cells spanning many generations and their genomes may differ, resulting in interpersonal and intra-tumor heterogeneity, which makes the study of cancer biomarkers difficult. Several publications have described prognostic and predictive biomarkers in different types of cancer as indicators of how a tumour is reacting to a therapy [48, 52]. By assessing tumour growth via biomarker expression, there is a high likelihood that personalised patient treatment may be possible [55, 65, 64]. However, it has been difficult to standardise the analytical methods for predictive biomarkers due to pragmatic issues such as the scoring system by using either immunohistochemistry (IHC) or Fluorescence immuno-localised samples. Quantitative assessments of samples by a pathologists visual scoring are subjective and vulnerable [132], while computational image analysis offers the potential to provide accurate and precise sample measurements [41].

To study cancer in a large and systematic manner, advanced techniques must be utilised. One of the most important techniques is conducting HCS experiments using automatic image analysis. High content high throughput screening is the screening of a large cell population with multi-coluor fluorescent microscopy, in combination with image analysis software and advanced statistical tools [67]. High throughput microscopy has benefited from the development of custom hardware, which includes auto-focusing and sample positioning, and from the development of image analysis software. In research, cell based high throughput technology is also widely used for target discovery and validation, as well as for epistasis analysis. Moreover, high content screening is appealing to biological and cancer researchers because it can define protein functions and dissect signalling pathways.

2.3 Imaging System

Basic concept of fluorescent confocal microscopy One of the most widely used techniques in cancer research is fluorescent confocal microscopy. The physical principles of fluorescent microscopy are based on the phenomenon of fluorescence and phosphorescence. The fluorescence phenomenon occurs when a molecule absorbs a photon with a specific wavelength and reemits light at another wavelength. Normally, a molecule stays in an energy state. When a specific photon is absorbed by the molecule, it gains sufficient energy to move to a higher energy state; however, when the molecule is in a higher energy state, it is less stable. Subsequently, the molecule emits a photon which has lower energy than the absorbed photon, as some energy is converted to heat. A molecule that can fluoresce is called a florophore [131], which can be detected using a microscopic system.

Since the birth of the first microscope more than five hundred years ago [122], a range of principles have been applied to design. The conventional design is transmitted light microscopy. Here, the physical principle behind image formation is the diffraction of light. Similar to transmitted light microscopy, phase contrast microscopy is based on the principle that light diffracted by a non-absorbing object has a phase shift of $\lambda/4$, where λ is the wavelength of the light. Both types of microscopy have a spatial resolution range of $10mm$ to $200\mu m$. Instead of using the reflection and absorption of light, fluorescence microscopes are based on the detection of photons that are emitted by molecules. The resolution of fluorescence microscopy is in the range from $200\mu m$ to $10nm$; within this range, cell structure is clearly visible. This type of microscopy is used in this research.

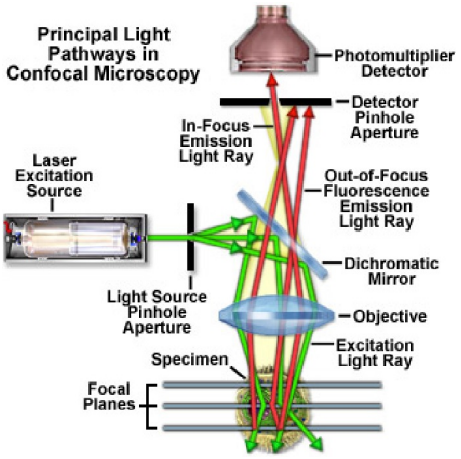


Figure 2.3: Principal light pathways in confocal microscopy [40].

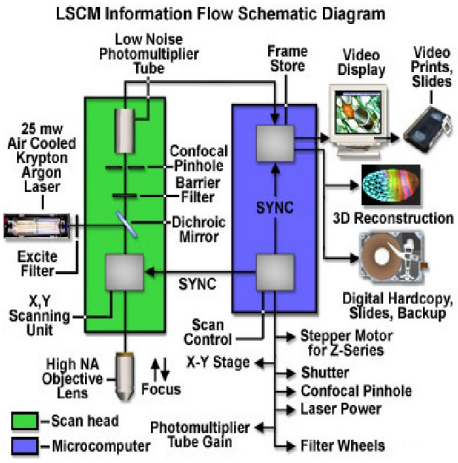


Figure 2.4: Information flow pathway in confocal microscopy [40].

A typical schema of a fluorescence confocal microscopy is shown in Fig. 2.4. The light source of the microscope is magnified by the lens and projected on to the Dichroic Mirror through a pinhole. Light travels through the epi-illumination objective to reach the sample (see Fig. 2.3). The light can excite the florochorome which becomes a secondary light source and it can travel through the Dichronic Mirror to reach the camera. Microscopic images are formed as a distribution of intensity in the camera. The pinhole eliminates out-of-focus light, as depicted in the Fig. 2.4.

The critical parameter in this system is the pinhole diameter [82]. A smaller pinhole affords higher spatial resolution, but also a higher noise level, as there is less energy. Further, due to the presence of a pinhole on the light source, the microscope illumination system has a low efficiency compared with a widefield microscope. Therefore, a stronger light source, such as laser light, is required in such a microscope. The settings of a confocal microscope influence image texture and optical aberrations, among other factors. The detailed fixation of confocal images is explained in Chapter 3.

The key advantage of a confocal microscope is that it can remove a significant amount of out-of-focus light. However, there is still some blurring at the edge of the point. The blurring on the edge of an object is defined by the point spread function (PSF). Deconvolution, which de-blurs the images, can be employed to reduce the effect of the PSF in microscopic images. Assuming that the image of an object is obtained by convoluting of the object image with a function, resulting in a blurred image, deconvolution attempts to invert this process in order to reconstruct the original object image from this blurred image. The challenge in deconvolution is to model the PSF. There are experimental, analytical and computational techniques to do this. In the experimental approach, images of small point-like objects are imaged and compared with the original object to obtain the PSF [106]. In the analytical method, the PSF is calculated using a classical diffraction based model [103]. In the computational approach, the PSF is approximated with different models and calculated according to these models. The computational cost of this last method is normally very high.

Besides blurring caused by the microscope point spread function, some blurring is caused by out-of-focus light. In addition, random noise that might be introduced by a non-uniform light source, affects the quality of image significantly. When the light source beam has different intensity, for instance, a very bright spot in the center of the light beam, the images obtained also have unbalanced intensity. Another source of noise is introduced by the optical elements in the system, whose performance can be affected by dust. As the images are recorded by electronic devices, these devices can also introduce noise in the images. The acquired images are also affected by the scattering of light within the tissue [148].

High content high throughput fluorescence microscopy High-throughput techniques refer to the automatic and fast analysis of a large numbers of compounds in biological, biochemical or biophysical assays. In these assays the interesting compounds or structures are rapidly identified [124]. The high-content technique refers to an operational manner in which a large number of measurements are collected for each specimen.A high-throughput high content fluorescence microscopy comprises of at least five steps. They are,

(1) Sample preparation. For example, in a high-throughput microscopy experiment, firstly, the samples of the tissue are prepared, including cell transfection and fluorescent labeling.

(2) Image acquisition. Then images of the sample are taken with a fluorescence microscope. The images can be multi-colour and three dimensional images.

(3) Data handling. As the amount of image data can be huge, special software and hardware are required to handle these data.

(4) Image analysis. Then the images of the sample are analysed by automated image analysis software, which is central to this approach. The goal of image analysis is to find biological objects of interest, segment these objects, and quantitatively analyse and classify them.

(5) Data mining. Last, the image data is modeled and analyzed by data mining techniques.

2.4 Image Analysis

Key challenges in the development of HCS are automatic image analysis and statistical interpretation of the images [137, 124]. The automatic analysis is typically a pipeline of analysis turned to a specific application. The main steps in a pipeline contains image processing (normalization, segmentation, tracking tracing etc), com-

putation of image features and machine learning for modeling and interpretation of data[1].

A list of contemporary popular image processing and analysis tools for life science research are listed in Table 2.1. Conventional image processing software normally provides many classic image processing and analysis algorithms, statistical tools and visualization functions that are used in an image processing pipeline. The functions and modules in current commercial and open source software packages are implemented for specific purposes in order to assist biologists in their experiments. The functions provided by software can be built into pipelines and allow users to design solutions specific to developer/user requirements. Open source software such as BioImageXD, Cellprofile and Image J also allow developers to write their own plugin functions. In contrast, the commercial software including Cellenger, dcilabs and Kalaimoscope, provides standard image analysis tools and cannot support modifications of their package. However, current available software cannot keep pace with all the demands of biological researchers and new innovative tools are required in order to test specific biological hypotheses. In addition, due to the growth of imaging data volume, many researchers are interested to mining the imaging data. In this case, if we want to use the images as data and quantify it, the imaging data has to be handled with precisions, while many commercial software are aim to provide tools for visual assessment and lack of systematic validations. The validation of algorithms are very important and it is also problem specific and in this thesis, all the processing algorithm are validated.

[1]Some application may also contain image registrations

Table 2.1: Image analysis software packages

Name	License	First Release	Implementation language
BiomImageXD	open source	2006	Python, C++
Cellenger	Definiens	2002	Cognition Network Language
CellProfiler	open source	2005	Python, Java
dcilabs	DCILabs	2001	IDL,Java
Image J	open source	1997	Java
Kalaimoscope	Transinsight	2003	N/A
Vaa3D	open source	2003	N/A
Konsanz Information Miner	open source	2003	N/A

2.4.1 Image Pre-processing

Microscopic images often contain intensity variations due to uneven illumination across the image, differences in image saturation, artifacts in imaging system and other variations due to the laboratory procedures. These effects can lead to problems in image segmentation and quantitative analysis. The objective of image preprocessing is to eliminate noise, normalize intensity values and make them suitable for the purpose of quantitative analysis. In order to estimate the uneven illumination, a typically means is to compute the mean or median intensity value at each pixel and then smooth it to generate a bias field, which can be used to correct the original image. Another typical step is to perform contrast stretching so that the intensity values has a predetermined distribution. The pre-processing step might change the pixel values which has an effect on the results of quantitative analysis and this needs to be taken into consideration when choosing the pre-processing methods.

2.4.2 Image Segmentation

Image analysis of HCS is often application driven. One of the most common objectives is cell and nuclei segmentation. The commonly used methods for cell segmentation and nuclei segmentation are Voronoi segmentation [81], model-based methods [36], seeded watershed [98, 94], active contour-based approaches [26, 114] and graphical

model segmentation [28]. The segmentation of cell and nuclei allows future statistics on a cell based analysis.

Most segmentation of cells requires two steps [81, 137]. The first step is to identify the cell nuclei and use is as seeds. The second step is to use the seed regions guide Voronoi segmentation and watershed segmentation. Thus, the automatic segmentation is nuclei is very important. Segmentation of nuclei is a process of labeling pixels in the image. Each pixel is labeled as either part of a nuclei or background. Following segmentation, it is possible to make morphological and quantitative measurements [122]. As in 2D images, there are overlapped nuclei. The main nuclei segmentation approaches which have been developed in previous research can be divided into bottom-up methods, active contours and level set methods, watershed methods and others, as described next.

Bottom up method The first step in the bottom-up segmentation method is the grouping of the adjacent pixels so that each group contains a statistically homogeneous region [51]. This local and coherent cluster is denoted a superpixel. One of the conventional ways of finding superpixel is using watershed transform of the image [51]. Then the superpixels are examined and merged together based on the similarities of some chosen features, or some particular rules in order to restore a full object [62, 94]. For instance, the average intensity of the superpixel can be a feature. A challenge in using this method comes from the similarity measurement or rules used to reconstruct the original object, because the common features between superpixels are not obvious. Finally, machine learning methods can be used to automatically group superpixels together.

Active contour and level sets Active contours and level sets are now classic methods and have been used widely in image segmentation. The basic idea in active contours (in 2D) is that an optimal curve is found which minimizes the total energy of the contour or a snake [84]. An energy-based active contour method is not able to deal with topological changes such as merging and splitting. Level sets overcomes this limitation and because topological changes often occur in cell analysis (e.g. mitosis) this approach has been used by many researchers to segment cells and nuclei. The usage of level sets can be divided into two groups; one is model-based, the other one is direct level sets. The model-based method evolves the level sets with the knowledge of the shape or size of the nuclei [112]. In applications where the objects are consistent in shape and size, it is a good method. Padfield, et al [120] used model-based level sets to segment cells in G2 and S phase4. Direct level sets do not require the knowledge of the model; it has also been widely used in the segmentation of cell nuclei [119, 112, 130].

Graph partitioning methods and learning based segmentation Graph-based segmentation algorithms usually construct and then partition a probability graph where the nodes correspond to image pixels or image regions and the edges are weighted and reflect the affinity between nodes. In order to compute the edge weights, local image features such as image intensity, spatial derivatives, maximum stable extremal regions [2] and texture etc. has to be used to construct the affinity graph.

Once the graph is computed, several algorithms can be used to partition the graph and segment pixels that have similar appearance into groups. One of the typical algorithm is graph cut. In graph cut, the seed pixels, node S and T, represent source and sink labels which are used to estimate foreground and background regions statistics [18, 19]. Al-kofahi [4] used graph cut to extract foreground information and detect

[2]Features (area, shape etc.) of regions computed using level sets.

seed points of nuclei by a distance map combined multi-scale Laplacian of Gaussian filtering. The seed points were then used to initiate another graph cuts incorporating the method of alpha expansion and graph coloring to reduce the computational complexity. Graph-based methods perform well when the regions to be segmented are coarsely labeled [137, 28, 18]. Similar to graph-based methods, learning based segmentation methods also require complex features associated with each pixel and then statistical classifiers are used to group pixels into each labels. The popular classifiers include random forest, decision tree models, support vector machines and so on.

Watershed methods The watershed method is based on an analogy from geology [110]. The image is assumed as a topographic surface with hills and valleys. It is assumed that basins exist in each valley bottom. If this image surface is flooded from these basins, the water will progressively flood the catchment basins. At the end of this flooding procedure, each minimum is completely surrounded by 'dams'. The set of dams obtained in this way corresponds to watersheds. This class of method is very sensitive to noise, as even if there is only a small minima, this method will find the 'dam' around it. As the watershed method can lead to over-segmentation of the object, a conventional usages of watershed is marker controlled watershed method [152]. A marker of the image is made as a first estimation of the nuclei. Instead of flooding water from the basin, it is flooded from the marker on object, so that the oversegmentation of objects can be avoided or reduced.

2.4.3 Quantitative Analysis

The quantitative analysis of microscopic images is mainly based on statistical learning and data mining techniques. As the information is analysed in different scales, the analysis can fall into three different categories: pixel-level analysis, cell-level analysis (including nuclei level analysis) and image-feature level analysis.

Pixels-level analysis In the case of pixel-level analysis, each pixel is considered as a data point and their properties are extracted from the images. One common objectives is to study the differences between groups of pixels that are derived from the images. For example, when both normal and cancerous phenotypes exist in the image, the pixels labeled as normal groups is compared with them labeled as cancerous. The comparison and study of the two groups might provide interesting insights. The methods that can be used with pixel level analysis include graph partitions, clustering and histogram-based segmentation methods etc. Most applications use pixel-level analysis when cell-level is not available. However, pixel-level analysis can not provide information about each individual cell, which is important for some experiment.

Cell-level analysis The cell-level analysis is based on success segmentation of cells in images. From the segmentation results, many geometric features and temporal parameters can be analyzed. The structure and dynamics of cell population can be caught and the cell cycle phase can be analyzed from the images [121]. In addition, the division rate of the cells can be estimated and form time-lapse images.

The phenotypes of cells are also very important measurements in experiments. The phenotypes of cells can be influenced by drugs, ectopic protein and other factors

[145, 126], such as their shape, size and volume. In the content of drug discovery, differences in phenotypes indicate the effect of drug. The automatic quantitative multidimensional measurements of individual cells have been developed by Zachary, et al [127]. In addition, there might be a strong link between phenotype and ectopically expressed protein in the cell [145]. This ectopically expressed protein might be a consequence of the presence of a cancer gene. An image-based multivariate method for classifying treated or untreated human cancer cells regarding to their phenotypes is developed [100]. The understanding of phenotypes will lead to a better understanding of cancer and cell behavior. To serve this purpose, quantitative morphological profiling methods was proposed to investigate the link between cell morphology and genes in a high throughput manner[10].

Another important measurement in biological analysis is the localization of proteins. The location of a protein can provide information about relevant metabolism within a cell [60]. Change of the location in time is a critical parameter in interpreting the behavior of the cell. There is a lot of research on the detection of protein localization [74, 37]. Huand and Murphy [73] have published a classification method which can detect the localization of protein with more than 90% accuracy.

Image features-level analysis In high content screening system and fluorescent microscopy images, image features can be computed directly from images or can be obtained from the sub-windows in images. These features can be derived from a single fluorescence channel or a combination of several channels. The computation of image features is important when more information is needed to be studied. For example, in order to analyze cell patterns, the image features related to cell patterns need to be included, such as rotation invariant or translation invariant features, Har-

alick texture, Zernike moment, object-based features, wavelet transforms and Fourier transforms, features from multiresolution subspaces and so on. These features allow us to have a better 'description' of the images, but not all of them are useful for every applications. In order to build a good model, meaningful and useful features has to be selected and identified. In practice, problem-specific features is normally designed for cancer biology problems to improve classification accuracies. For instance, the overlap between a biomarker and a nuclear can be used as a feature in study of micro-tubules.

Feature selection is important when we analyze image based on image features. The feature selection step is strongly related to the machine learning models that will use to analyze these features. In general, the selection of features can be conducted using filter methods, sequential feature selection, principal component analysis (PCA) [79], independent components analysis (ICA) [75], Fisher linear discriminant [135], Isomap [13], neural network [69] and so on.

2.4.4 Related Work in Machine Learning and Computer Vision

Machine learning in medical imaging is an emerging field where algorithms or machines are designed that can 'learn' from data so that the algorithm or machine can 'interpret', 'inference' and 'predict' data. The machine learning methods are heavily based on statistical learning, where statisticians try to extract information, patterns and trends from data and try to understand what the data says [113, 16]. In high content screening analysis, most machine learning problems can be categorized as supervised learning and unsupervised learning problems. When the outcome measurements of data is available, supervised learning methods can be applied, otherwise, unsupervised learning is more applicable.

Supervised Learning Supervised learning aim to build a model of the distribution of class labels so that it can perform classification and regression on new data [113, 16, 137]. Supervised learning can be used to classify cells with different morphology. By compute cell width, circumference and convexity these morphological features can be used to create a classifier. For example, intensity and nuclear spot features were computed and used to classify and profile nuclear phenotypes. Classification can also be used to study different patterns in cell images and recognize different objects. The most popular classification models are the support vector machine [56], decision tree model [97, 43, 139], neural network and graphical models etc [86]. Among all the supervised classification models, support vector machine and random forest are two of the most popular tools and they are explained in details (in Chapter 5).

Unsupervised Learning Unsupervised learning is a powerful data mining techniques where the data is unlabeled and there is no assumptions about relationship between data points. The unsupervised learning algorithms seeks to discover and characterize the distribution of data or its hidden relationships. The common methods in unsupervised learning include k-means [80, 3], hierarchical clustering [78, 5], etc. Many supervised learning algorithm with modification can also be applied in unsupervised fashion, because they are all based on statistics and information theories. The application of unsupervised learning can help finding how the centrosome duplication is affected by a number of different drugs by construct hierarchical clustering tree [29, 125, 137].

2.5 Tissue Image Analysis

In order to study signalling pathway in tumour and discover cancer biomarkers of clinical use. It will reduce labour, reduce bias and improve efficiency if automatic HCS can be used for imaging tissue specimens. However, due to the complexity in cell morphology in tissue samples, it has been challenging to automate the quantification.

The number of publications in the research field of tissue image analysis is relatively smaller than the number of publications in cell segmentation and tracking algorithm for cultured cells. There are a few publications on tissue images. Most of them used patch based analysis [29, 32, 92, 4, 89]. Because sometimes, it is impossible to segment cell or cell nuclei from these images. In addition, the current methods have not considered automatic tissue quality in their study.

2.6 Conclusion

Even though there has been many research on the automatic and quantitative analysis of microscopic images for high-throughput high content application, there is not a robust tool for tissue images, due to the complexity of biological images. There are commercially available software packages, for example, CellProfiller and Cellomics. They cannot meet all the needs of biological researchers. The microscopic image analysis for high throughput purpose is still a challenge for researchers.

In the previous research, most of the samples are cultured cells. As cultured cells are in a different environment from *in vivo* cells, their behaviours under drugs or other stimuli might also be very different. There is a strong need to investigate cells from

tissues besides cultured cells. Due to scattering of tissue and variety of phenotype in tissue, it is still a challenge.

Since the samples in the images are different, the previous method can be modified and combined to achieve the best results. For example, Nath [116] proposed a cell segmentation algorithm using Level Sets with coupled constrains and graph-vertex colouring methods and classify image pixels into either cell or extra-cellular category using colour-texture features extracted at the local neighbourhood of each pixel. When building an application, the choice of methods are normally task specific and problem specific. In addition, when using learning based method, the features are normally subjected to the samples. It is important to choose proper methods for the applications.

Chapter 3

Pre-processing of Tissue Microarray Images

3.1 Introduction

In this doctoral thesis, multi-channel fluorescent confocal microscopy techniques are used to study cancer biopsies. The need is to identify and quantify cellular and sub-cellular components from these images. The manual analysis of images is labour intensive, time consuming and contains human bias. Therefore, computer software solutions which can automatically segment and analyse images are desired as assistive tools in modern laboratories.

However, the imaging data typically contains many artefacts from the imaging system and the imaging data can need to be normalised before automatic analysis can be applied. Previous research has suggested that due to the limitation of optical systems in microscopes, the microscopic images are always corrupted [141, 122]. When a microscope is used as an instrument for quantitative purposes, minor artefacts in the optical system can cause large errors in quantification results and invalidate the analysis. For instance, a previous study suggested that without the correction of uneven illumination, the successful detection of cells decreased by 35% [141].

In order to make fluorescent confocal microscope suitable for quantitative analysis, it is very important to perform image pre-processing. The pre-processing of images aims to correct uneven illumination and normalise the data so that they can be quantified. Due to the limitation of the hardware, it is a very challenging to purely use computational methods to eliminate bias and improve the accuracy of the imaging data. This chapter starts with a thorough explanation of a typical experimental procedure, followed to acquire data in our work, followed by an analysis of systematic errors. We then present the methods we have used to pre-process the imaging data.

3.2 Explanation of imaging data

Ewing's sarcoma is a rare form of tumour affecting mainly children and young adults. The five year survival rate[1] is around 75.8% for children younger than 15 years and 49.0% for age between 15 to 19 years old [142]. Most Ewing's sarcoma tumours are caused by a chromosomal translocation[2]. The chromosomal translocation leads to the formation of a EWS/FLI fusion protein between the protein EWS and an ETS [147]. This fusion protein is believed to act as an aberrant transcription factor[3] involved in regulating cell growth and differentiation. However, the mechanics which introduce tumourgenesis in cells are still unclear. Therefore, there has been considerable interest, including by our collaborators in biological sciences, to study the signalling pathways in Ewing sarcoma biopsy.

Ewing sarcoma tissue microarrays (TMA) were derived from the from the Institute of Pathology, University of Valencia; the Institute of Pathology, University Medical Center, Desseldorf and the Department of Pathology, Leiden University Medical

[1] the percentage of patient survived until five years after being diagnosed with Ewing's sarcoma.

[2] this chromosomal translocation is an abnormal event when two chromosomes incorrectly recombine, such as the EWSR1 gene on chromosome 22 fuses to the FLI1 gene on chromosome 11 [147].

[3] transcription factor is a protein that binds to a specific DNA sequence, controlling the rate of transcription of genetic information from DNA to messenger RND

Centre in collaboration with EuroBoNet [54]. Importation of anonymised tissue was approved under UK HTA licence and local ethical review to the Sarcoma component of the Oxford Research Biobank. From these biopses, one or several tissue cores around 6 mm diameter were taken and embedded in paraffin block. The formalin fixed and paraffin embedded pre-treatment biopsies were subsequently processed so that nuclei were labeled with DAPI and membrane was immune-labeled with mouse anti-CD99.

The images of the sample were acquired using a confocal microscope Olympus Fluoview fv1000. Each image has a size of 2048×2048 pixels with a horizontal and vertical dimension of $211\mu m \times 211\mu m$. All images were taken with a 60x oil objective. for each specimen, between 1 and 6 images were acquired. The limitation of the imaging system is explained in details as they affects the accuracy of the quantitative results. After the image was acquired, the pathologist used the Olympus Fluoview ver 2.0b viewer to convert the imaging data to 'tiff' image format for further image analysis. As shown in Fig. 3.1, the fluorescent signals are acquired using four different channels separately[4].

A low and high threshold were determined while the microscope operator took the image. However, it was often the case that the images were obtained by several different biologist and the consistency in setting the low and high threshold by eye was hard to maintain. In order to find the best low threshold, it was found to be better to take the images where the predominent peak[5] in the low intensity range and then to normalize their values using computational method. The high threshold was

[4]there might be detection of other channels. explain a bit more.

[5]the predominent peak in the intensity histogram of the image, which is indicating the majority of the pixels are at certain value. since the background pixels is normally has a larger population in low intensity, and they are always the largest groups of pixels, it is then can be sure that the information about cells were not lost.

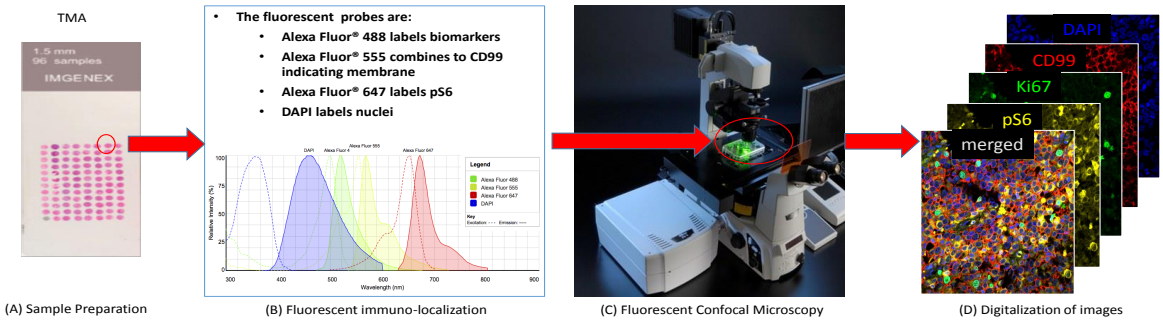


Figure 3.1: The flow chart showing the procedures of acquire confocal images.

also determined by the biologist by eye-balling where the pixels need to be saturated. When the image is heavily over saturated, the cell nuclei will looks like a constant intensity (non-textured white). The information about cellular pattern is hard to be observed. The channels were then assigned with four different colors so that their relative location was easier to observe.

3.2.1 Accuracy and precision in quantitative analysis

In order to quantify the images and extract useful information from the fluorescent confocal microscopic images, it is important to have a thorough understanding of the limitations in the system, from sample preparation to digital image formation. It is also important to understand the source of error in the data so that the imaging data is analysed in an accurate way.

Variations in sample preparation. The quantitative analysis of TMA images aims to quantify variations in the TMAs and discover new information from them. However, the construction and storage of TMAs can introduce loss of antigenicity on paraffin slides and directly affect the quality of immunohistochemistry staining and any related quantitative analysis [156, 53, 49]. Previous research has discovered that the following factors may have a considerable impact on the quality of immunohistochemistry: storage time, oxidation, hydrolysis, tissue processing time, light exposure, fixation time, and the type of fixative used [150]. Some of the common problems are explained in details next. Firstly, during the fixation process, the amount of fluorescent probes binding to the sample is difficult to control even using the standardized protocols. Secondly, the fluorophore emission can be influenced by other molecules in the sample, resulting in 'quenching' of the fluorescence from the excited fluorophore. Thirdly, after the sample has been exposed to light, the 'brightness' of the sample can also fade with time, which is a phenomenun known as photobleaching. Finally, due to insufficient removal of the non-bound fluorescent probe and auto-fluorescent of a sample, there may be high background fluorescence.

Another common challenge in clinical tissue microarray quality control is the mixture of tumor tissue, stroma tissue and non-tumor cells in a tumor biopsy. The existence of other type of cells in tumour biopsy can result in failure of the image analysis software and introduce significant noise in quantitative analysis.

Uneven lighting and vignetting. One of the commonly observed artifacts of an imaging system is uneven lighting and the vignetting effect (shown in Fig. 3.2. In Fig. 3.2(a), a gradual darkening of the image from one side of the image to the other can be observed, which might be a result of incorrect centering of the lamp filament

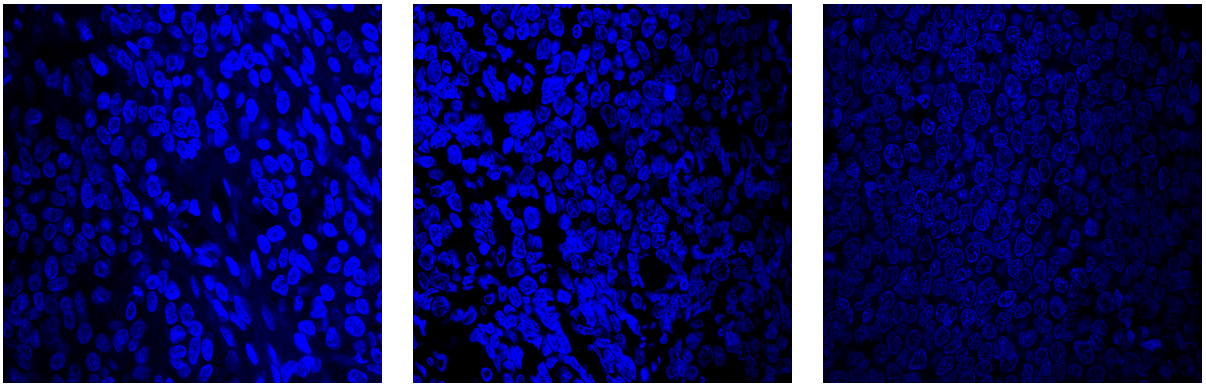


Figure 3.2: Example of TMAs Images. (a) shows the tissue nuclei (DAPI stained) image is brighter at right side of the image than it is on the left side. (b) shows another example where the left side of the image is brighter than the right side. (c) shows the darker corners on the image which is known as vignetting. This error occurs when the camera shutter speed is too fast to allow the film to respond or it might be a result of field curvature of the camera lens.

[117]. In an ideal illumination condition known as Köhler illumination [2], the vignetting effect as shown in Fig. 3.2 is due to the radiance crossing the objective lens that decreases proportional to the $cosine^4$ of the angle at which light impinges on the sensory array [2]. However, in practice, other factors can contribute to the vignette, such as lens impurities, the angular sensitivity of the digital sensor and its response function [144, 61]. This presents challenges to correct the images using computational methods sometimes referred to as retrospective correction.

Variations in offset and saturation. After the biologist has captured the images, it is impossible to recover the signals that are not captured by the camera. The normalization of the imaging data is performed based on the character of the experiment, but it is impossible to obtain a total clean data set. As shown in Fig. 3.3, the saturation of the image may lead to a lost of texture in the nucleus. In an experiment where we are interested in quantifying A, B and comparing their intensity values, the results of it will contains errors from this saturation. The images were normalized

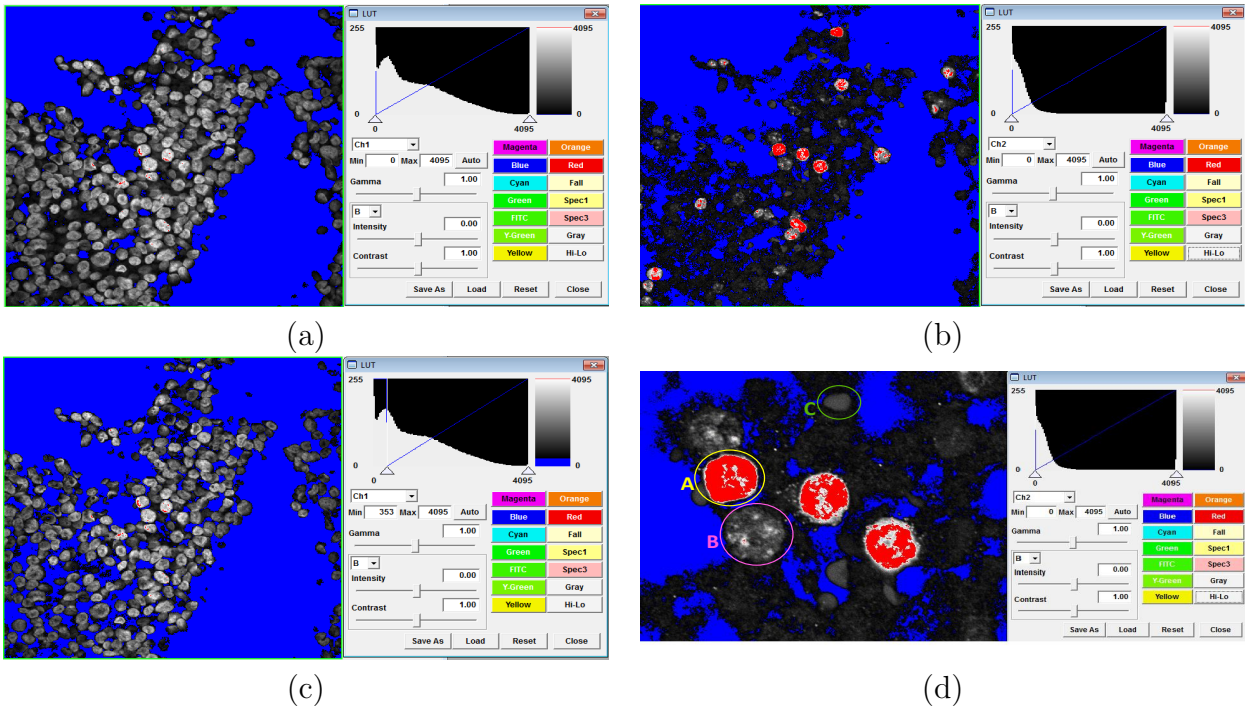


Figure 3.3: An example of the digital image obtained from the microscope using Olympus Fluoview software. (a) the DAPI channel is shown in hi-lo model, where the blue pixels indicate their value being 0, while the red pixels indicate their value being 4095 and the values in between them are showing in a gray scale. (b) an example showing the effect on the image, when the low threshold were changed by biologists. (c) the cell nucleus A (marked by yellow circle) is shown an over-saturated nucleus while another nucleus B (marked in magenta circle) is lower in intensity and showing a pattern of chromosome. The nucleus marked by green circle is a different type of cell showing different pattern. The reason that nucleus A is much brighter than nucleus B might related to their cell phrase. (d) examples of saturation and lost of nuclei pattern because of it.

by the AQUA threshold method [24]. However when the uneven-light and vignetting effect exists, their impact should be taken into consideration.

3.3 Retrospective Illumination Correction of Confocal Images

3.3.1 Literature Review

The problem of uneven illumination has been addressed by several researchers in previous publications [128, 9, 93]. There are two conventional way of correction the images. First, when the bright field images and black field images are available, the illumination function can be calculated using equation 3.1.

$$BiologicalObject = \frac{DigitalImage - BlackfieldDigitalImage}{BrightfiledDigitalImage - BlackfieldDigitalImage} \quad (3.1)$$

This is known as prospective illumination correction. In equation 3.1, $BiologicalObject$ is the true light from the biological object. A $BrightfieldDigitalImage$ can be obtained in experiments when the a digital image is taken without a biological sample, but with an sample of homogeneous fluorescent illumination [33, 134]. A $BlackfieldDigitalImage$ is obtained when a digital image is taken without a light source, but keep the setting of the digital camera unchanged. In previous research, the $BlackfieldDigitalImages$ term is often neglected [128, 45, 154, 158]. The image can be corrected using equation 3.1.

In practice, bright field images are not always available, so that a retrospective correction of illumination is needed. This means that it is necessary to correct the bias in illumination using mathematical models. Previous research has suggested a few common methods of retrospective correction of the microscopic images [93, 95, 59, 9, 151, 123]. The general mathematical models behind these methods assume that, at each pixel x, there is an biased illumination function $v(x)$, which is either multiplicative to the acquired image and an additive term $z(x)$ which is related

to camera offset and thermal sources (shown in equation 3.2). The true image that we want to recover $I(x)$ related to the observed image $I_0(x)$ shown by:

$$I(x) = \frac{I_0(x) - z(x)}{v(x)} \tag{3.2}$$

In previous literature, the methods developed for retrospective correction of illumination can be categorised into four main classes of method. The performance of these classes of method performance is not as good as prospective correction, but their advantage is that they can be used for a single image. In many studies, only one single image is available for analysis.

The four classes are:

(1) the illumination is corrected based on parametric models. The model is normally built based on theoretical and physical properties of the light distribution, such as the cosine-fourth law of the intensity decreasing in the field[66, 154, 68, 25]. However, parametric models do not incorporate the shape changes on the vignetting curve due to the impurities of lens, misplacement of light source and dust etc.

(2) methods that model the vignetting effect as an additive low frequency signal which can be extracted from the image [93, 95, 59, 9]. These methods work better when the vignetting effect is very strong and a prior knowledge of the radius of vignetting effect is required. In addition, some methods are based on morphological filtering. This approach has been shown to give the best performance compared to Gaussian filtering and Polynomial fitting methods in [9].

(3) methods are based on automatic or semi-automatic segmentation of background, where the background is estimated by fitting a quadratic function, B-spline or other planes to the segmentation results.

(4) methods takes the advantages when several images are acquired use the same microscope with the same settings and the background is estimated based on the characteristics of local statistics (pixel intensity distributions over overlapping images or mosaics). More advanced method proposed by Smith et al. [141] need at least 10 images to produce a good estimation of the vignetting effect. The performance of multi-image based methods can achieve as good accuracy as a prospective method. The validation of these methods are based on reference images.

3.3.2 Methodology

In this thesis, the images were acquired at different times; the set up of the confocal microscope might have changed during an experiment; and there is no record of which images were taken at any one time. The methods suitable for solving this problem is retrospective correction of a single image. As shown in the Fig. 3.3, the blue regions in the images are assumed to be background and their values were set to zeros by the biologist, which make it hard to use the method based on the segmentation of background. The drifting of the intensity values due to illumination bias can only be observed when a biological sample is observed at the same time. In addition, there is no reference images[6] and it is not possible to validate the algorithm use conventional peak signal to noise ratio (PSNR).

In order to solve this image correction problem, a few reasonable assumptions are made. The first assumption is that the uneven illumination field can be characterised

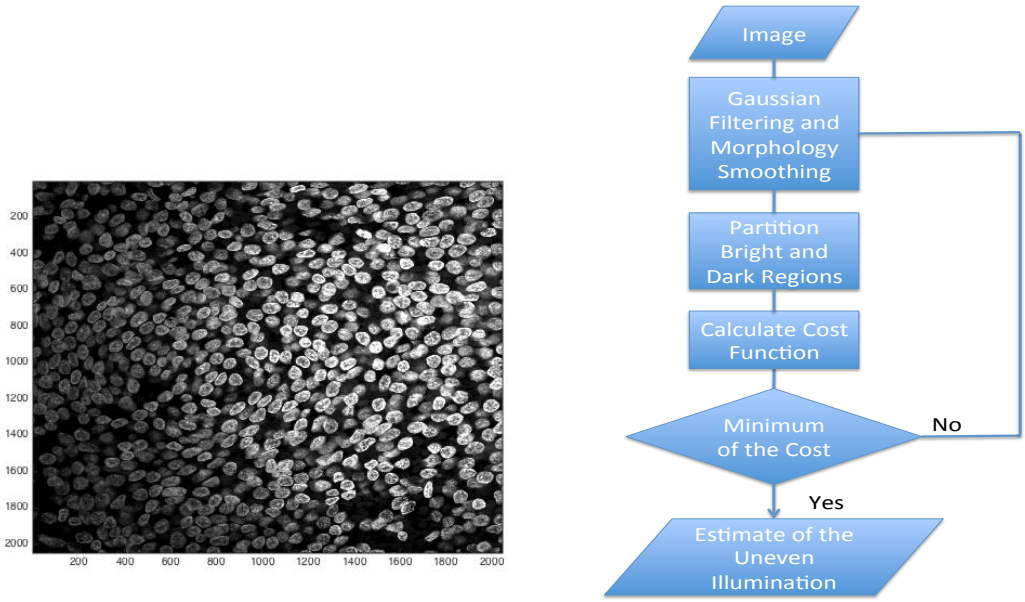
[6]There is not an image taken without the specimen in place.

by the low frequency component of an images [93, 9]. Secondly, the frequency of this low frequency component is relatively stable across all the images. As shown in Fig. 3.4, an uneven illumination field has only one bright peak [93, 9, 128, 141]. All the images were taken at the same spatial resolution and there was only one peak in the uneven illumination. In a single image, the intensity distributions of the same sample were the same across image regions. The difference between regional intensity statistics was mainly due to uneven illumination. As shown in equation 3.3, the correction of the image can be considered as a two step process. The first step estimates the multiplicative term $v(x)$ and the second step estimates the additive term $\frac{z(x)}{v(x)}$. The bright region object before correction and after correction has the same intensity range. The intensity range is adjusted so the region before and after correction has a similar intensity range. The saturation adjustment is assumed to be equivalent to correcting the image with the term $\frac{z(x)}{v(x)}$.

$$I(x) = \frac{I_0(x) - z(x)}{v(x)} = \frac{I_0(x)}{v(x)} - \frac{z(x)}{v(x)} \tag{3.3}$$

Overview An original algorithm was implemented to correct the uneven illumination in the images. The method employs a Gaussian filter and regional intensity statistics to estimate the uneven lighting. The idea of correcting uneven illumination using a Gaussian filter is not new, but there is very little literature validating the Gaussian filter quantitatively when reference images do not exist. The challenge is to build a method that can correct for vignetting based on only one single image and does not require reference images.

The flowchart of the algorithm is shown in Fig. (3.4 (b)), the digital images were filtered by a low pass Gaussian filter and morphology operators and an estimate of



(a) (b)

Figure 3.4: An example of the retrospective illumination of confocal microscopic images. (a) the observed image acquired by the confocal microscope. (b) An overview of the correction algorithm

the uneven illumination v obtained. In order to assess the accuracy of the estimate of true uneven illumination of the image, a cost function was defined by statistical analysis of the intensity distributions of the bright and dark regions.

Gaussian filter The isotropic Gaussian filter is a low-pass filter with a 2D Gaussian kernel shown in equation 3.4. Here, σ is the standard deviation of the filter.

$$G(x,y) = \frac{1}{2\pi\sigma^2} e^{-\frac{x^2+y^2}{2\sigma^2}} \tag{3.4}$$

The low frequency component in the image is obtained by convolving the observed image I_o with this Gaussian kernel (equation 3.5).

$$B = I_o(x,y) * G(x,y); \tag{3.5}$$

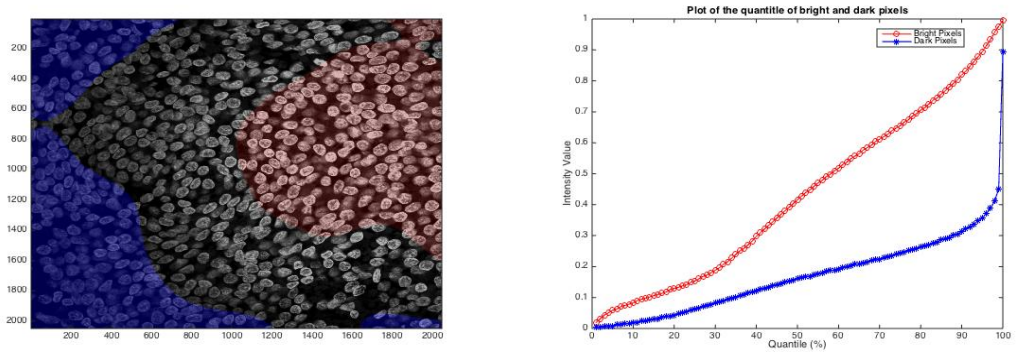
B is an estimate of the uneven illumination which correspond to the low frequency

component in the image. Obviously, B is determined by the Gaussian kernel and the optimal parameters of Gaussian kernel need to be determined.

The most important parameters of Gaussian kernel is the size of the kernel[7] and σ. It is important in the experiment to find the optimal kernel size and σ, which are large enough to eliminate high frequency components, but also can capture the uneven illumination. The ideal kernel size and σ give the minimum value of the cost function that defined in 3.6.

Bright and dark regions In order to find the bright region and the dark region, a low pass Gaussian filter and morphological smoothing is used to segment bright regions and dark regions. The image is firstly convolved with a Gaussian filter which has standard deviation σ and a low frequency signal B is obtained. The intensity values of the B is smoothed by morphological filtering. The morphological filter employed is a 300 pixels 'disk' element to dilate and erode the illumination field. In B, the region containing intensity values smaller than 25% quantile of the intensity (in B) is defined as the dark region (group x). The region containing intensity values higher than 75% quantile is defined as the bright region (group y). The pixels in bright regions and dark regions are treated as two groups of data points. Due to the limitation of image saturation and offset, the pixels with zero intensity value and the ones with maximum intensity values need to be removed from the analysis. As a result, only non-zero and non-saturated pixels in bright and dark regions are included in group x and y respectively. As shown in Fig. 3.5, the bright region and the dark region are computed and the intensity values of the pixels in these two regions were compared with each other using the empirical quantile functions.

[7]The full width at half maximum (FWHM) of Gaussian is $2\sqrt{2\ln 2}\sigma$



(a) (b)

Figure 3.5: An example of the bright and dark region computed by the algorithm. (a) The bright region is labelled using red color and the dark region is labelled with blue color. (b) The quantile of the intensity values of pixels in bright region and dark region.

Cost function The cost function is defined as the differences of intensity distributions between the pixels from the bright region and the pixels from the dark region. The (empirical) quantile function Q of the distributions of the intensity values are computed for group x and group y respectively . F_n is the distribution function of the intensity x, $X_n(1), X_n(2), \ldots, X_n(n)$ are the order statistics of the sample.

$$Q_n(p) = inf\,\{x : F_n(x) \geq p\} = X_n(i) \tag{3.6}$$

where i is chosen such that $\frac{i-1}{n} < p \leq \frac{i}{n}$ and $X_n(1) \leq X_n(2) \leq, \ldots, \leq X_n(n)$.
The mean square differences of the two quantile functions is used as the cost function.

$$cost = \frac{\sum_{p=0.01}^{p=1} (Q_x(p) - Q_y(p))^2}{100}; \tag{3.7}$$

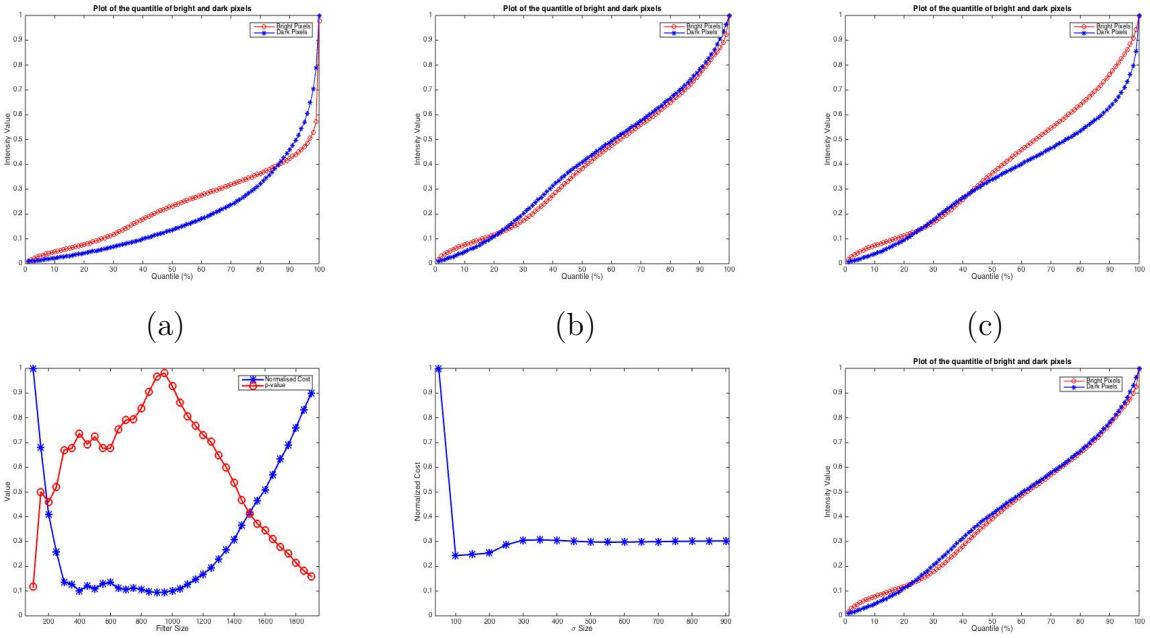
The student's t-test is also used to confirm the finding that by changing the size of the Gaussian filter and σ, the statistical difference between dark and bright regions can be minimised. The null hypothesis is that the data in vectors x (pixels from the bright region) and y (pixels from the dark region) comes from independent random samples from normal distributions with equal means and equal (unknown)

variances.The p-value of the t-test is also shown in Fig. 3.6(d) to confirm the similarity between dark and bright regions.

$$t = \frac{\overline{x} - \overline{y}}{\sqrt{\frac{{s_x}^2}{n} + \frac{{s_y}^2}{n}}} \tag{3.8}$$

Optimisation The size of the Gaussian filter is optimised against the cost function. A set of parameters were obtained from the experiment. The size of the Gaussian filter has a strong effect on the calculated cost. So the size of the filter is optimised first and σ is set to a quater the size of the filter. As shown in Fig. 3.6 (a)-(b), the differences between dark and bright region were computed and a minimum value is obtained when the filter size is set to 900×900 pixels. The σ value is optimised by setting the filter size to 900×900, as shown in Fig. 3.6 (e).

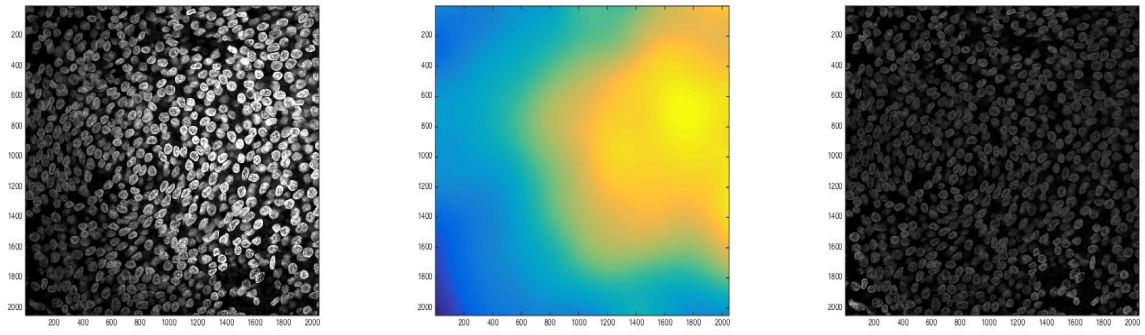
After the optimised Gaussian filter size and σ are obtained from the optimisation step, the uneven illumination is computed use equation 3.5, as shown in Fig. 3.7. The digital image is corrected by dividing the uneven illumination estimate v (equation 3.3). However, using this method, the function z is difficult to estimate. According to equation 3.3, the true image contains the image we obtained in Fig. 3.7 (c) and an additive term $\frac{z}{v}$. We have already obtained v so that the effect of z can be observed in the histogram of the intensity of the image. The existence of $\frac{z}{v}$ is observed in the intensity distribution of the image. Instead of estimating z directly, the intensity of the image is adjusted so that the brightness of the nuclei is in the same range before correction and after.



(a) (b) (c)

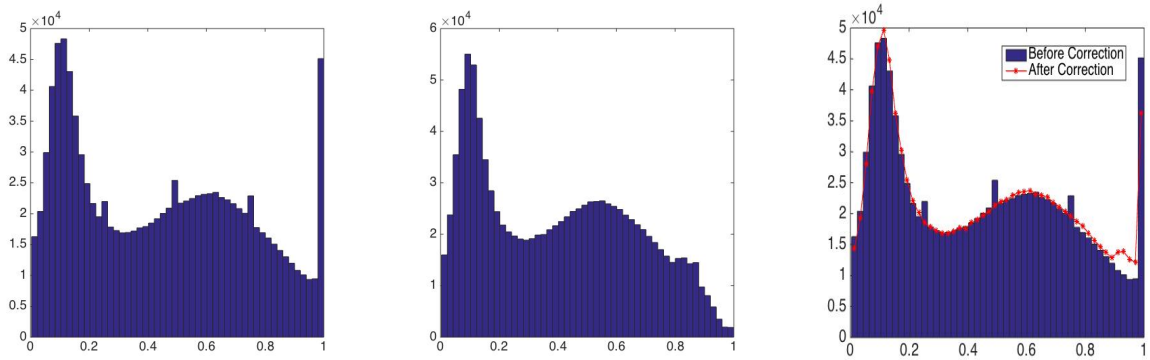
(d) (e) (f)

Figure 3.6: An example of the optimisation of the Gaussian filter size and σ. (a) An example of empirical quantiles of dark and bright regions after filtered with a 100×100 pixel Gaussian filter. (b) An example of empirical quantiles of dark and bright regions when applied a 500×500 pixel Gaussian filter. (c) An example of empirical quantiles of dark and bright regions when applied a 1900×1900 pixel Gaussian filter. (d) The value of the cost function against filter size (σ is set to half of the filter size). (e) The value of the cost function against σ (filter size is 900 pixels). (f) The quantiles of dark and bright pixels when filtered by the optimised filter size (900 pixels) and σ (150 pixels)



(a) (b) (c)

Figure 3.7: An example of the uneven illumination obtained from the method and the results after the image is corrected by this uneven illumination function.



(a) (b) (c)

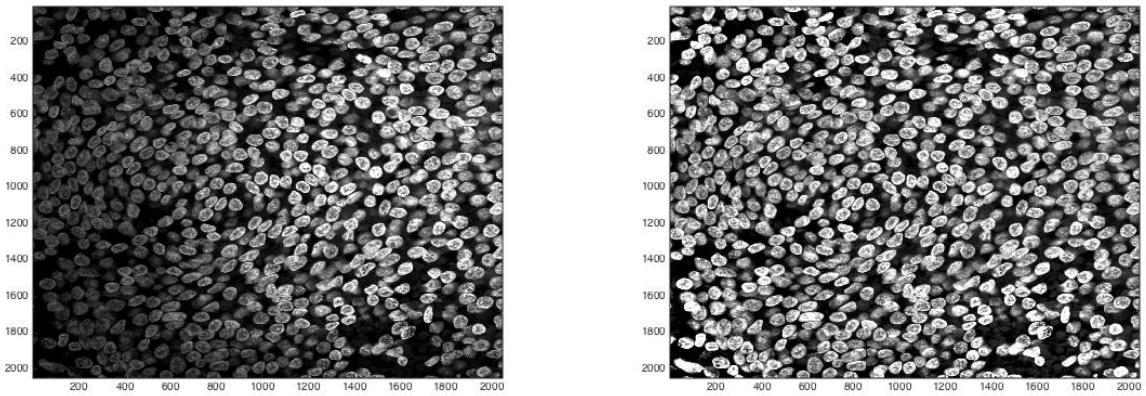
Figure 3.8: An example of the saturation correction. (a) The histogram of the original image. (b) The histogram of image after filtering, the number of saturated pixels has decreased comparing with (a). (c) The histogram of pixels after saturation correction is plot against the original intensity distribution.

3.3.3 Image Intensity Saturation Correction

After correcting by users the estimated illumination function, the intensity of the image shifted towards lower intensity. This shift can be observed in the bright region of the image. The aim of the saturation correction is to bring the intensity of bright regions back to its value before it is divided by the illumination. As shown in Fig. 3.8, the shift of the intensity histogram is observed in the experiment and it needs to be corrected. The pixels that has been saturated in the original images were selected and their value in the the corrected image were computed. The saturation of the image needs to be adjusted. The value of the lowest intensity values of the saturated pixels were compute and it is used as the new saturation threshold T. In this experiment T is in the range $[0, 1]$. The range of intensity values is adjusted by this threshold.

$$I_n(x) = \left\{ \begin{array}{ll} 1, & I(x) \geq T \\ \frac{I(x)}{T}, & I(x) < T \end{array} \right\} \tag{3.9}$$

By visual assessment it can be seen that the retrospective correction has produced an even illuminated image, as shown in Fig. 3.9. We validate the method in the next



(a) (b)

Figure 3.9: An example showing the image before and after correction. (a) the original image. (b) the image after correction.

section.

3.3.4 Validation

The algorithm has been validated on 58 randomly chosen images from the data set and the same Gaussian filter was applied to the images. Due to the lack of reference images, it is impossible to compute the conventional signal to noise ratio to validate the method. The images were validated by visually examining pre and post correction images.

The mean square difference (MSD) of quantile functions between bright and dark region was calculated. The mean of MSD of the 58 images was 2.12 and the standard deviation was 3.53. This means that for most of the image, the MSD was very small but there were a few images with a large difference. An good example is shown in Fig. 3.10(a) and (b). The worst example is shown in the Fig. 3.10 (c) and (d), the image before correction and after correction is shown in the image. However, due to the spatial variation of the cells, the dark regions are mainly background and the sample doesn't have the same intensity distribution in the bright and dark regions. This leads to a higher value MSD. But by comparison the corrected image, the dark corner of the image is corrected by the algorithm and a more even illumination can be observed. This result suggests that when the samples has different intensity distributions in bright and dark regions, MSD can not indicate well how the image has been corrected. In this case where there is no reference image, a visual examination is necessary.

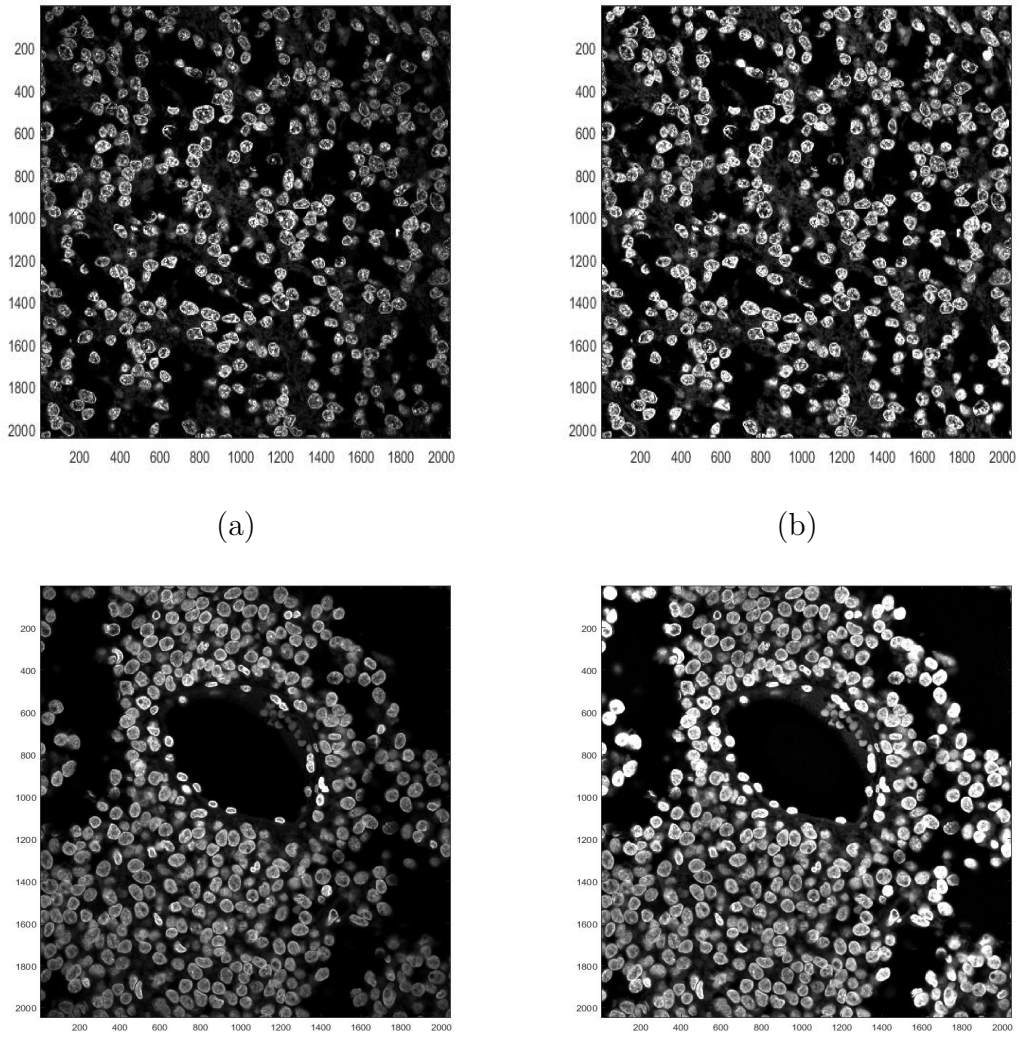


Figure 3.10: The visual examination of images. The images at the left hand side are the original images. The images at the right hand side are corrected images.

3.4 Illumination Correction and AQUA Threshold

In order to make sure that the imaging data is corrected in an unbiasedly, all the images in this experiment were pre-processed using the algorithm developed in the previous section. The Gaussian filter employed, as defined in equation 3.4 was the same for all the images and no human bias is introduced in this process. AQUA threshold is a common adjustment of the fluorescence image is developed in [24]. Only the targeted signal is selected and the AQUA threshold determines the predominant peak of the image and sets a threshold slightly larger than this predominant peak. By apply this adaptive threshold method, the images are normalised. The lower threshold, which is arbitrarily set by the pathologist is normalised. As shown in Fig. 3.3 (c), by changing the lower threshold of the image to the predominant peak of the intensity histogram, the amount of background pixels eliminated is normalised. By automatically perform this adaptive threshold on all the images obtained by the pathologist, the difference of the camera offset is eliminated. If the image does not have a predominant peak, there might be information loss of lower intensity information when the image is taken and it is not possible to use this adaptive threshold method to normalise it.

3.5 Summary

Source of errors deriving from the acquisition protocol used in this project have been discussed in this chapter and a careful preprocessing and normalisation of imaging data proposed to be performed prior to further analysis. The preprocessing of imaging data is specific to this dataset and practical issues are taken into consideration. However, the visual assessment of illumination correction might be needed for certain images and all of the images are automatically corrected.

Chapter 4

Image Segmentation of Tissue Micro-array Images

4.1 Introduction

Automatic segmentation of the cell nuclei and cytoplasm in tissue sections is an important step prior to many quantitative analysis in high content screening (HCS) experiments. The segmentation of individual cells and nuclei from tissue biopsies allows us to take measurements from single cells and also provides the potential to study cell populations and their interaction in great detail. In addition, by studying each individual cell automatically and unbiasedly, it enables us to carry out systematic study of tumour heterogeneity[1] which may contribute to better understanding of resistance to cancer therapies [6, 107, 104, 143]. In order to study heterogeneity of cells and inter-cell communications, it is essential to establish single cell measurements from HCS images.

Previous work on automatic image analysis of HCS images described in the literature, mostly focuses on the analysis of images from cultured cells and not tissue samples. A robust algorithm to segment cell nuclei and cytoplasm from microscopic

[1]Phenotypic and functional differences between tumours of the same type in different patients (inter-tumour heterogeneity), and between cells within the same tumour (intra-tumour heterogeneity) and they are very often observed in previous research

images of fluorescent stained tissue samples in a fully automatic manner is highly desirable. In addition, the validation of a fully automatic algorithm is very important to the accuracy of quantitative analysis and none of the previous methods have not been well-validated against this dataset. This chapter addresses challenges in cell segmentation and develops an original automatic segmentation algorithm.

4.2 Collaborations

The work described in this chapter has been published in journal form [23], and was a collaboration with a pathologist and statistician. In this paper, the laboratory work was carried out by Claudia Bühnemann in the Dunn School of Pathology, University of Oxford. The random survival forest analysis was carried out by Simon Li as part of his doctoral research, from the Institute of Biomedical Engineering, University of Oxford. The automatic image segmentation was carried out by myself as part of this doctoral research and the method developed is explained in details in this chapter. The main contribution of the author is the implementation and validation of the OxPathBio image segmentation software.

4.3 Literature Review

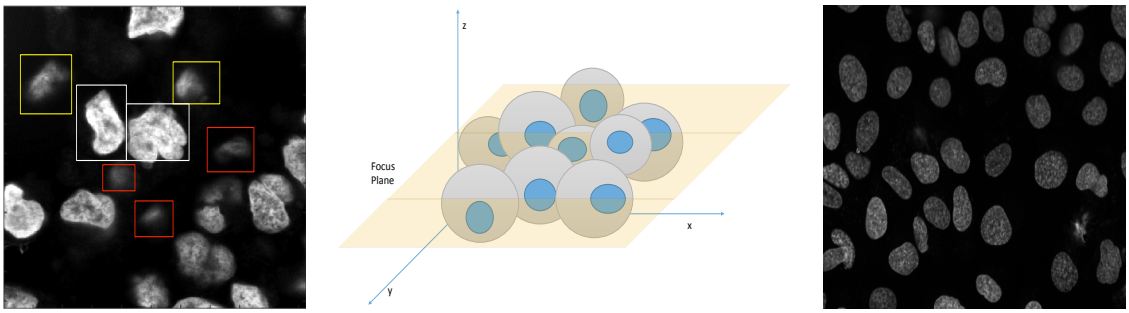
Cell nucleus segmentation is a well-studied problem. Previous methods include thresholding [30, 118], watershed [98, 94], deformable models (active contour and level set) [15], morphology-based methods [96] and combinations of them [1, 118, 27]. One of the key challenges in cell segmentation is segmentation of the cell nuclei when they are overlapped with each other. This problem has been addressed using several approaches which were shape models and curvature analysis [155, 27]. For instance,

Arsenal et al [7] developed a model-based nucleus segmentation algorithm for overlapped cells. In these methods, an estimate of the average radius of the nuclei is often required. In addition, automatic [42, 115] and semi-automatic methods [31] have been proposed. Learning based methods learns the size, shape, intensity distributions, or other features from a training image dataset, and use the derived classifier to segment super pixels or pixels.

However, the shapes of the nuclei in tissue samples vary a lot due to disease, age, gene mutations and other factors [149], which prevents the use of curvature analysis and shape models. To address this challenge in our work, an original algorithm has been designed and is implemented to split clumps of nuclei. The algorithm is based on the assumption that when nuclei overlap with each other, the size of the clump of nuclei is proportional to the nuclei clump within it.

Most previous methods have been developed and validated on cultured cell lines [108]. In addition, many of the previous methods have been developed on a relatively small dataset, around tens of images from several cell lines. However, in our applications we are interested in segmenting over a thousand images from over a hundred patients. In this case, a segmentation technique that is robust to all kinds of cell morphology change of is needed.

Cultured cells are often processed to form a flat layer of cells [101]. By contrast, cells in tissue and cell cores grow in three dimensions. As shown in Fig. 4.1, the cells may be orientated in any direction. Compared with cells in cultured cells, cells in cell cores and tissue have more irregular shapes. By taking an image of the focus plane, the objects in the image can be in-focus nuclei, or out-of-focus nuclei. By purely looking at the image, it is not possible to tell if the centre of a cell is on the



(a) (b) (c)

Figure 4.1: Compare of cellular images of cell core and cultured cell. (a) An example of tissue image. (b) Cells grow 3 dimensionally can have different orientations. (c) This is an example from a well-know cultured cell dataset [35]

focus plane, or lower or higher than it (Fig. 4.1). In addition, only looking at the DAPI channel, it is also hard for pathologist to tell if the nuclei in the middle is a big nucleus or several smaller overlapping ones. Because the underlying structures of cell population are very different, the models developed for cultured cells is not always useful for tissue cells. It is still challenging to segment cells in tissue samples. We developed an original cell segmentation algorithm to segment cell nuclei and cell cytoplasm in this tissue sample.

4.4 Overview of the Segmentation Algorithm

In order to segment cells from the images, a model that can adapt to morphology changes is highly desirable. The segmentation algorithm I have developed consists of several steps and has several components to segment both cell nuclei and cytoplasm from an image. Figure 4.2 summarises the pipeline I have developed. The DAPI channel, which highlights cell nuclei is first pre-processed by the method explained in Chapter 3 Section 3.3. Nuclei locations are obtained by a level set model, and this location indicator is later used to produce a background marker by a watershed transform. With the guidance of a background marker, another level set is used to

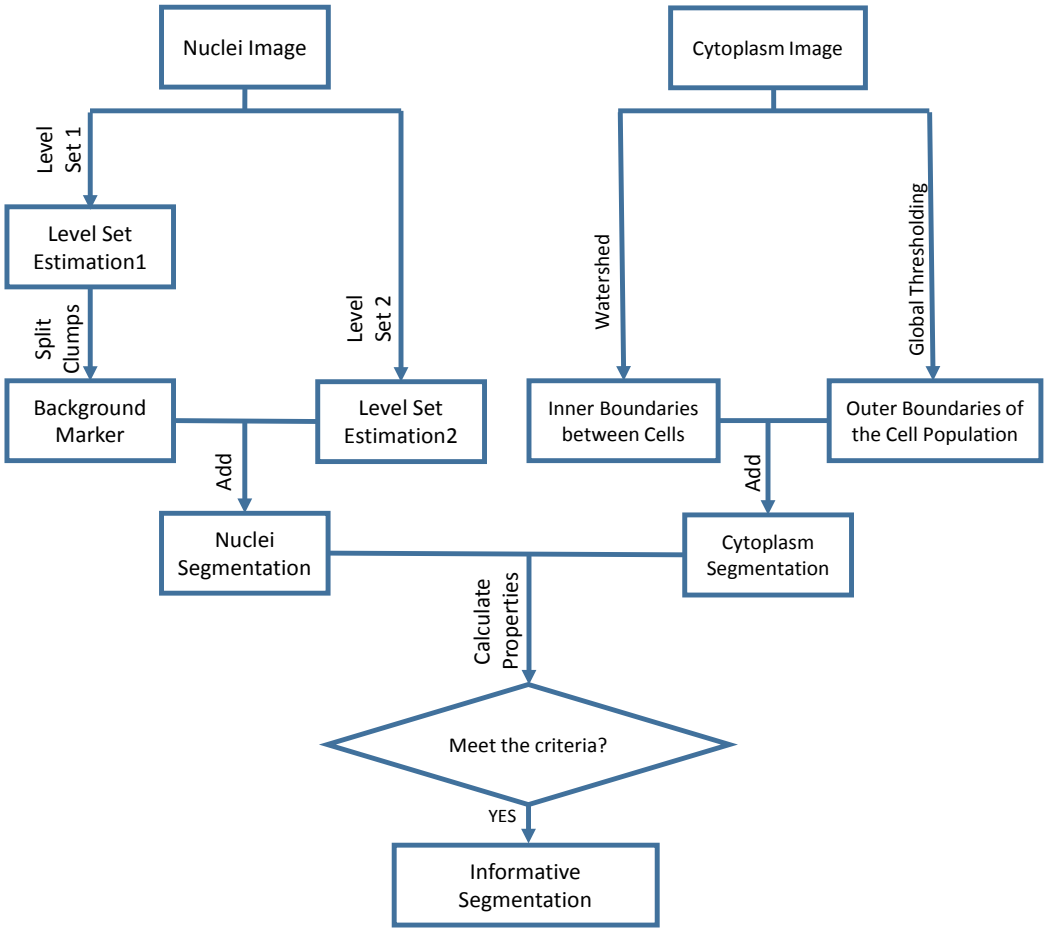


Figure 4.2: Overview of the image segmentation algorithm

precisely find the boundary of the cell precisely. The results of segmentation are then used as an indicator of the nuclei location. The segmentation of cytoplasm requires finding the outer boundary of a cell group and the boundaries between cells. The objects identified and segmented by this method in a DAPI channel is then processed together with its corresponding cytoplasm segmentation. Measurements taken from the object and cytoplasm are used to decide if the detected object is an informative cell with an informative nucleus and cytoplasm.

4.5 Segmentation Algorithm

In this section, we detail the key steps of the developed algorithm.

4.5.1 Segmentation of Nuclei

An original nuclei split algorithm has been implemented which is based on multiple level sets and a machine learning algorithm. In the following text, the level set method is explained and the details of each components in the segmentation pipeline is explained. In addition, the link between each components and the choice of parameters are explained.

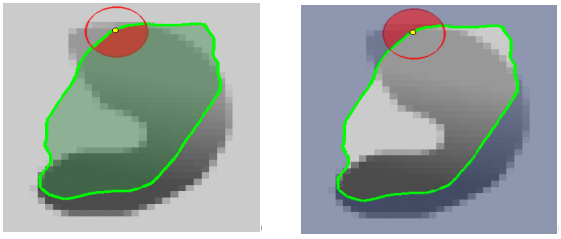
The segmentation of nuclei contains two main steps. The first step detects the cell nuclei and produce an estimate of nuclei location and size. The second step uses a level set to find the boundaries of the nuclei. In the first step, a hybrid geodesic region-based level set serves to estimate location and shape of individual nuclei [90]. This estimation is corrected by using a Watershed tessellation. Based on the corrected estimation, another level set curve is initialised and the objects boundaries are obtained.

4.5.2 Level Set

Hybrid Active Contour The combination of a region and a geodesic level set was used in this work. The energy definition underpinning the level set formulation was:

$$E = \oint_{C(s)} \int_{x\in\Omega} (I_\chi(x,s) - \mu_l(s))^2 dx + \int_{x\in\Omega} (I_\chi(x,s) - v_l(s))^2 dx ds \qquad (4.1)$$

Here, $C(s)$ represents points on the curve (green curve in Fig. 4.3). $\mu_l(s)$ is the mean intensity of the neighbour points in the interior region (the red region in Fig. 4.3



(a) (b)

Figure 4.3: An example showing the local interior and local exterior regions [90]. (a) Local interior. (b) Local Exteriror.

(a)). $v_l(s)$ is the mean intensity of the neighbour points in the exterior region (the red region in Fig. 4.3(b)). χ represents the local neighbour, which evaluates the neighbour points by 1 and other points by 0. The mean interior is obtained by equation below:

$$\mu_l(s) = \frac{\int_{x\in\Omega} I_\chi(x,s)dA}{\int_{x\in\Omega} \chi(x,s)dA} \qquad v_l(s) = \frac{\int_{x\in\bar{\Omega}} I_\chi(x,s)dA}{\int_{x\in\bar{\Omega}} \chi(x,s)dA} \tag{4.2}$$

In order to evolve the curve, the curve flow function is defined here as a function of s. The change in the curve with iteration number can be written as:

$$\begin{aligned} C_t(s) = &\int_{x\in\Omega} (I_\chi(x,s) - \mu_l(s))^2 k\vec{N} + \int_{x\in\Omega} (I_\chi(x,s) - v_l(s))^2 k\vec{N} \\ &- \oint_{C(s)} (\mu_l(r) - v_l(r))(2I_\chi(x,s) - \mu_l(r) - v_l(r))dr\vec{N} \end{aligned} \tag{4.3}$$

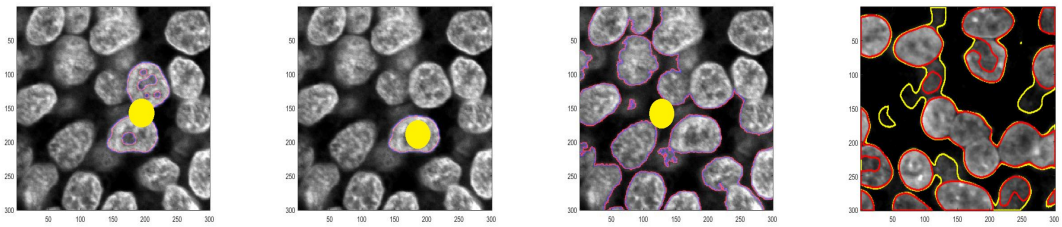
In this equation, r and s is from 0 to the length of curve C. To derive C, each point s on C is calculated and r traverses the curve at point s. k is the curvature of the curve at point s. The equation provides a direction and speed for the curve to travel. When a curve is near a boundary, the first two terms are smaller. So the flow of curve is slower. The third term provides the direction in which the curve travels.

Inner cut off intensity In nuclei, there is normally a strong intensity texture. In a texture weighted level set model [83], the texture can be defined in the energy function. In this work, nucleus texture did not show an obvious patterns, and the intensity inhomogeneity often led to a failed segmentation. To solve this problem, an adjustment of intensity is made. For each pixel $I_{original}$ in the image, the equation below is used to calculate a new intensity I_{new} for the pixel.

$$I_{new} = {I_{orignal}}^2 \sqrt{\frac{1}{1 + (\frac{I_{orignial}}{A})^2}}; \tag{4.4}$$

Here, A is the inner cut off intensity. The intensities which are higher than A are reduced. The intensity which are smaller than A are not changed very much. The choice of cut off intensity is crucial as it determines the behaviours of the level set. This parameter has to be selected empirically for a given image. In all our experiments, it is set to the value of the first quartile of the initial intensity histogram.

Parameters in Level Set In this implementation, the growth and behaviour of the contour is controlled by the parameters in the model. The parameters include defining the size of local neighborhood (χ in equation 4.1), defining the maximum number of iterations and setting the size of each step in the speed function. The choice of these parameters affects how fast the model converges, but does not affect the segmentation results significantly. The segmentation results of the model are mainly affected by its initialization and inner cutoff intensity. It is important to compute the best initialization of the level set. In this work, the maximun iteration number of evolution was set to 30. The radius of the local region was set to 16 pixels.



(a) (b) (c) (d)

Figure 4.4: An example showing the initialisation of level set affects the segmentation results. The contour of the yellow regions indicates its initialisation. The blue line is the line indicating outer side of the curve while the magenta line indicate the inner side of the curve. (a) The initialisation include part of two nuclei and background pixels. (b) The initialisation locates at the centre of the nucleus. (c) The initialisation located at the background captured background pixels.(d)An example showing the results of level set initiated with different threshold values. The red line indicates the results from an initial of higher intensity values, and the yellow lines indicates the results from a lower initial values.

Initialization of Level Set Good initialization of the level set is very important. A bad initialization can lead to failure to capture objects of interest. For example, three initializations were manually made on an image and then passed to the level set model. Results are shown in Fig. 4.4. As shown in Fig. 4.4 (b), when initializing the active contour at the right location, the contour can segment a nucleus. In Fig. 4.4 (a), the initial contour is located in between two nuclei. The contour grows from the initial location and split into multiple contours.

The initialization intensity is also important. Different intensity threshold of initializations were used and result to different segmentation results obtained. As show in Fig. 4.4 (d), by setting the initialization of the level set with different thresholds, the segmentation results are significantly different. When using a higher intensity value for initialization, the evolution of the model contains brighter pixels. While setting a lower value for initialization results in a much larger contour that contains more pixels. In addition, the lower value in initialization enables the contour to include more low intensity pixels relative to a higher values. For example, the lower

intensity components such as out-of-focus nuclei can be excluded when setting the initial value to a higher number.

4.5.3 Cell Detection

In order to segment nuclei in a fully automatic manner, it is necessary to develop an automatic seeding algorithm. Due to the existence of background noise and out-of-focus nuclei, images often contain nuclei as shown in Fig. 4.1 (a). It is unclear that the pixels marked by red and yellow boxes are noise, in-focus-nuclei or out-of-focus nuclei. However, the pixels in the yellow box formed a texture and the intensity of them seems strong, so it can be assumed that there are nuclei in the yellow boxes. In order to analyse them in future, we name all of these positions as cell point candidates (CPC). In the red boxes, the pixels are low in intensity, there are not obvious patterns and they are very small in size. It is highly possible that they are noise. We are interested in detecting the CPC from this image.

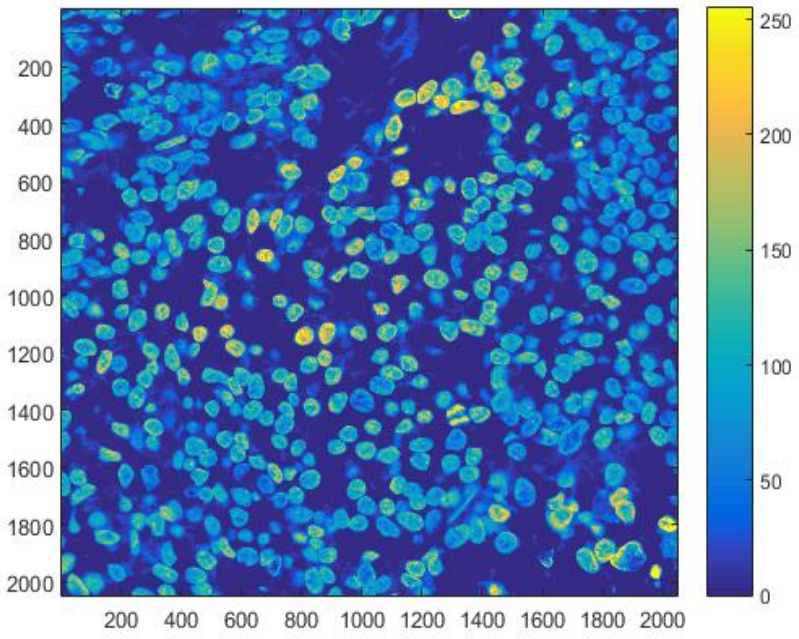
Multiple Level Set: By initiating a level set with a set of different intensity thresholds, a series of level set estimates are computed and superimposed as shown in Fig. 4.5. By taking the regions where the pixels appear very frequently, a binary map is produced and this mask is then processed to identify individual nuclei. In detail, the initial intensity threshold is computed use $\alpha \times T$, where T is the Otsu's threshold of the image and α is in an empirically determined range $[0.3, 2]$. For example (shown in Fig. 4.5), after adding all the level set results together, the in-focus nuclei appear more often than the out-of-focus nuclei. The frequency that each pixel appears in the series of level set segmentations are used as a feature to detect the in-focus-nuclei and the regions with frequency stable segmentation results are considered as CPC. A binary map is produced containing all the potential cell points in it. In practice, the

results of this binary map was found to be very close to that of a single level set where the threshold of intensity is $0.8 \times T$ and it was chosen to reduce the computation time.

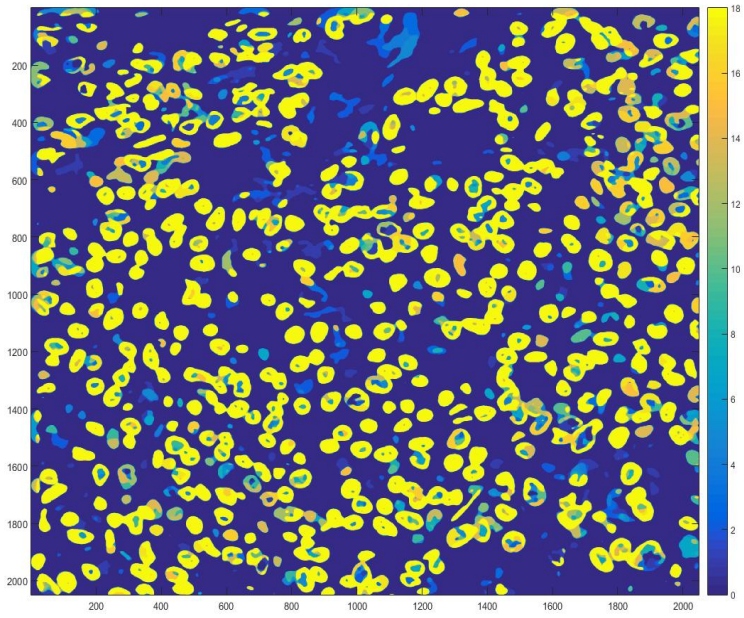
Split Clumps Often the binary map contains clumps of nuclei. An original clump splitting algorithm was used to separate each nuclei. The clump split method is an iterative classification and segmentation process. The classifier needs to decide which region to split and the associated method to split them. In this design, because the 'actual' size of the nuclei is assumed unknown, the area of a region relative to all the other regions in the binary map is used as a feature. By splitting the connected nuclei, the average size of the regions become smaller and this feature has to be calculated again to identify the clumps that are not previously detected. The clumps that are newly discovered by this step need to be segmented. By iterating this classification and segmentation process, the average area of the nuclei is gradually found and the nuclei are segmented from clumps (shown Fig. 4.6).

The segmentation of individual nucleus is achieve using a watershed based method. In order to overcome the under-segmentation and over-segmentation problem with classical watershed segmentation (shown in Fig. 4.7), a parameter H is used to remove local valleys that is smaller than H. If H is too small, the binary map is over-segmented, as shown in Fig. 4.7. If H is too large, the binary map is under-segmented. Due to the variations of size and shape, the 'ideal' H depends upon them. In addition, H can be a number or several numbers in a certain range. Because of this reason, the minimum number of H is adopted. In certain cases where H does not exist, these regions are not segmented.

As shown in Fig. 4.6(a), a classification model is needed to determine how to

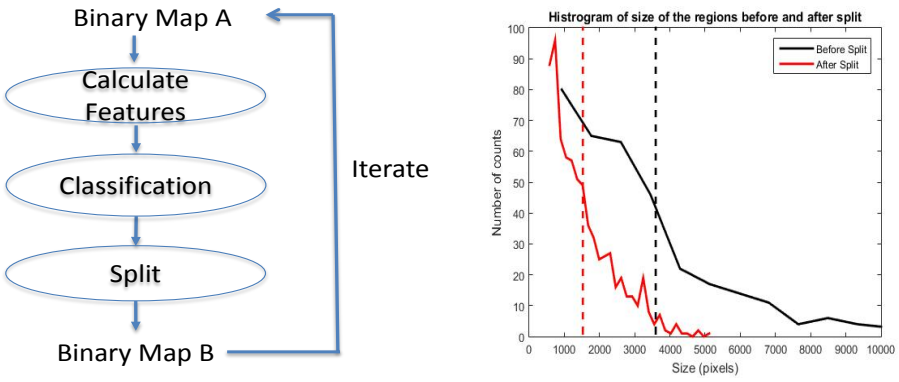


(a)



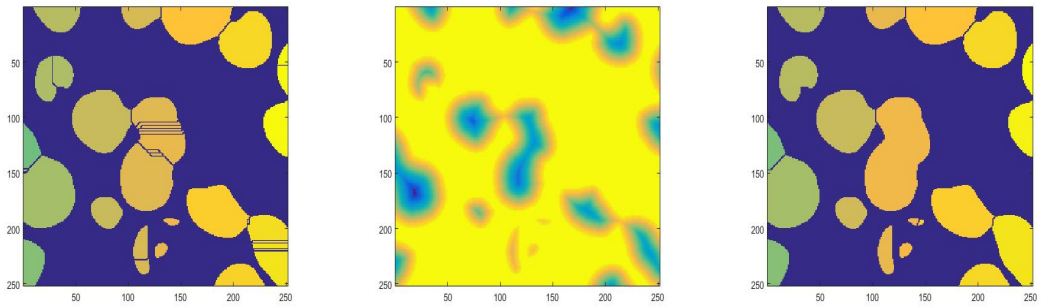
(b)

Figure 4.5: An example showing the results of level sets initiated with a set intensities. (a) the original image shown in a parula color range, so that it is easy to see the background in the image, which is difficult in a gray color range.(b) An example showing the results after adding 20 level sets together, the nuclei that are in focus appear in all 20 segmentations.



(a) (b)

Figure 4.6: Split Clumps. (a) The schema of splitting algorithm. (b) The histogram of clumps sizes after several iterations.



(a) (b) (c)

Figure 4.7: An example showing the over segmentation of nuclei. (a) an example of over segmentation of nuclei. (b) change of parameter H can smooth the distance map in watershed. (c) after tuning H, the over-segmentations are removed.

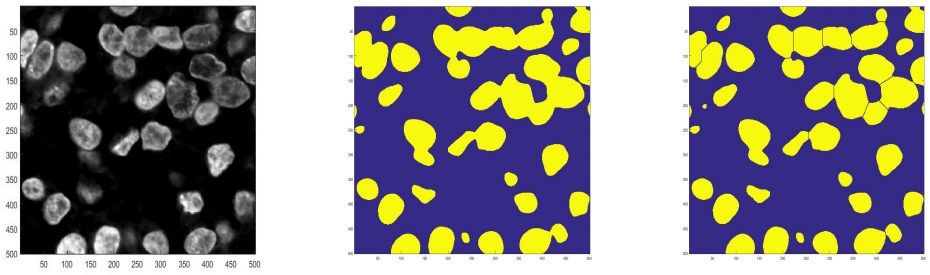
Class	Area	Solidity
Class 1	from 100 to 1.2S	or from 0.9 to 1
Class 2	from 1.2S to 2S	and from 0 to 0.9
Class 3	from 2S to 4S	and from 0 to 0.9
Class 4	from 4S to inf	and from 0 to 0.9

Table 4.1: The classification model try to classify the clumpy into one of these 4 classes.

split the cell. The classification model in this work, were manually designed with an associated method of segmenting the cells. The classes are defined using the following ranges (see Table 4.1):

Here, S is the average area of all regions in the binary map. Depending on which class and which round of iteration, the region is split, and an H is associated with this region. In the reported experiment, 3 iterations are adopted. As shown in Fig. 4.6 (a), the algorithm begins by calculating the area and solidity of the regions in the binary map A. The regions are classified in to four groups. For the regions belong to Class 1, the algorithm does not split it. For regions in Class 2, they are split with a particular H. For the regions belong to Class 3 and Class 4, because their size are in a larger range, 3 recurrent of this algorithm are performed on these regions. The results from Class 1, Class 2, Class 3 and Class 4 are then added together to form the results of first iteration, shown as a binary map B. The next iteration starts by calculating the mean area of regions in it.

In this work, 10 images were randomly used in the reported experiment, and H was determined empirically based on these 10 images. Further optimization with more images might give better results. However, with the current value of H, very good results are achieved.



(a) (b) (c)

Figure 4.8: An example showing the the segmentation of clumps of nuclei. (a) The original image in gray color map. (b) The initialisation locates at the centre of the nucleus. (c) The initialisation located at the background captured background pixels.(d) An example showing the results of level set initiated with different threshold values. The red line indicates the results from an initial of higher intensity values, and the yellow lines indicates the results from a lower initial values.

Cell Point Candidate Each region in the segmentation results are then considered as the cell point candidates. As shown in the example in Fig. 4.8, the image of nuclei are segmented using a level set and a binary map is obtained. By using the split clump algorithm, the nuclei that form clumps are partitioned. All the regions in this Fig. 4.8(c) binary map is a cell point candidate. These cell point candidate are then used to segment the cytoplasm of the nuclei and they will be evaluated together with their cytoplasm.

4.5.4 Segmentation of the cytoplasm

An example of a cytoplasmic image (CD99) is given in Fig. 4.9. In this image, the red colour is the CD99 staining of the cell membrane. The boundary between cells are not obvious.

Explanation of the algorithm The segmentation of the cytoplasm is based on computing the inner and outer boundaries of a cell (Fig. 4.10). The outer boundary

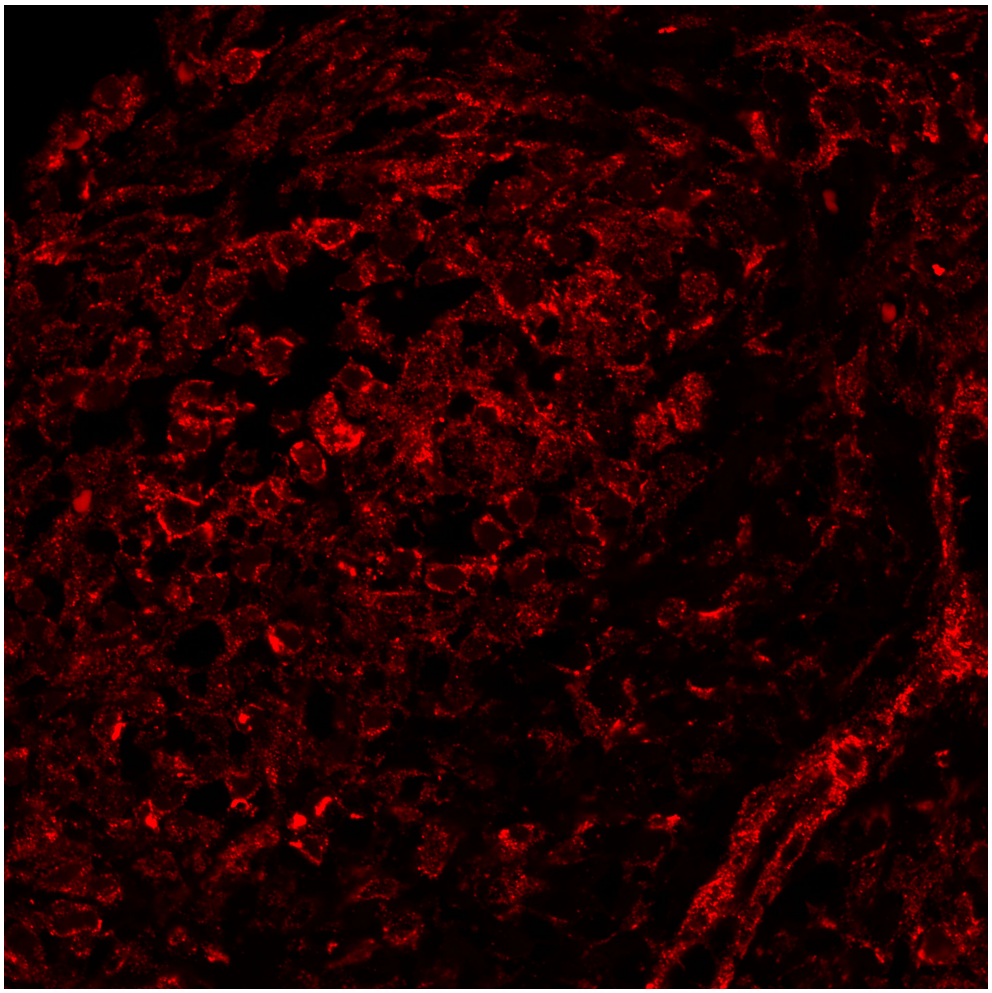


Figure 4.9: An example to show the CD99 staining of the cell membrane.

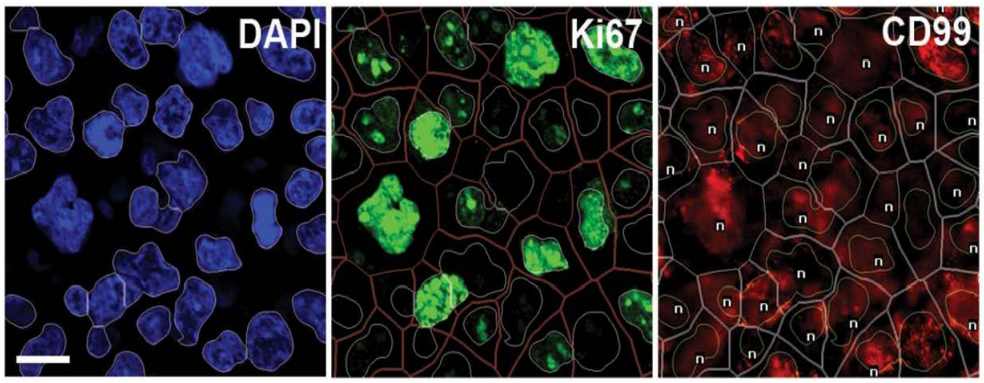


Figure 4.10: An example to show the segmentation of cell. The figure at the left shows the segmentation of cell nuclei. The figure in the middle shows the segmentation of cell nuclei and cytoplasm can be mapped on Ki67 biomarker channel. A quantitative measurement can be made to Ki67 based on segmentation results. The figure at right, shows the segmentation of cytoplasm based on equal distance between nuclei.

is computed using Otsu's threshold of the image and the regions that are greater than 3000 pixels are defined as CD99 positive regions. This size was determined by pathologists. The boundary of this binary map is the outer boundary of the cell population in the image.

The inner boundary is the boundary between cells, which are computed based on equidistant between neighbouring cell point candidate . After candidate cell points are obtained, the equal distance between neighbouring candidate cell points is. A watershed tessellation is used to find find the equal distance line in this map and partition the space into cells. This line is the inner boundaries between cells. The inner boundary and outer boundary are mapped to the cytoplasm image, and a watershed tessellation segmentation of cytoplasm is obtained.

4.5.5 Discussion about cell cytoplasm segmentation

The method is used because the unstable staining of CD99, which is caused by low CD99 expression, poor fixation or non-uniform staining and the membranes of cells is

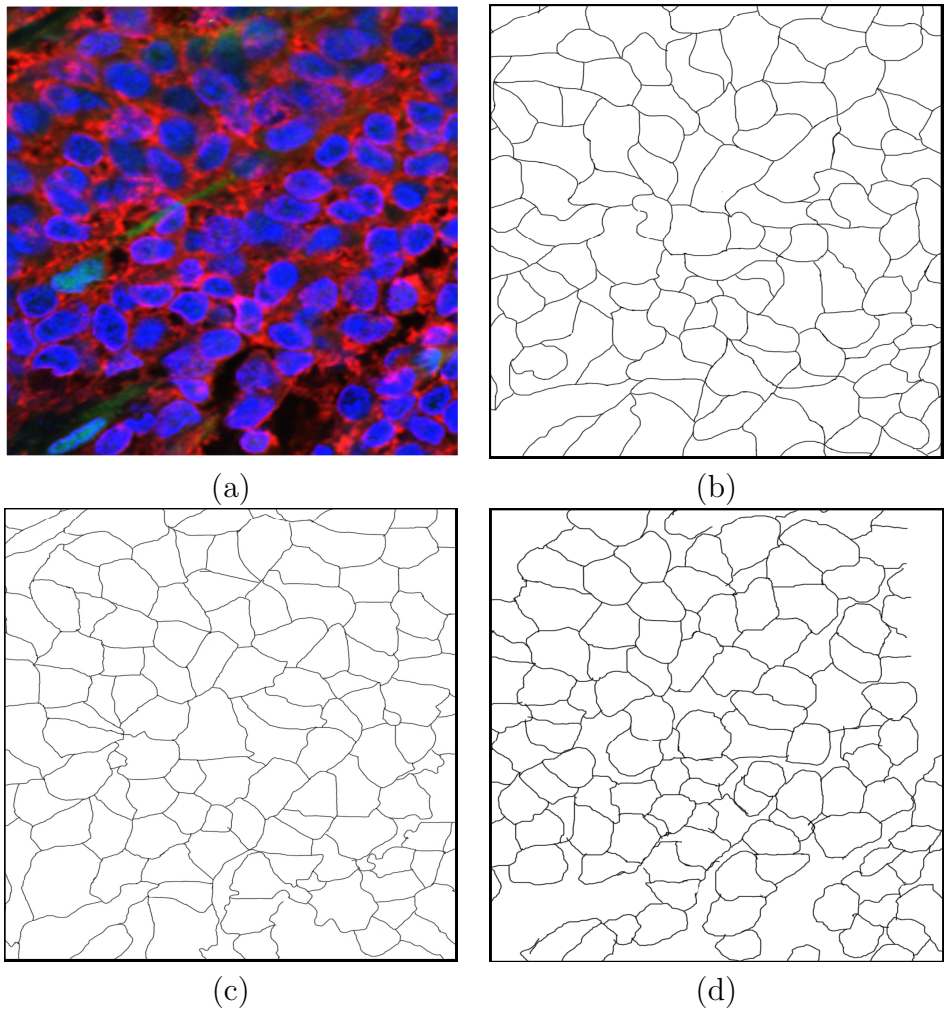


Figure 4.11: An example to show the difference of cytoplasm segmentation among experts.(a) the original tissue images. (b) manual segmentation of (a) by expert 1. (c) manual segmentation of (a) by expert 2.

not always detectable. This section demonstrate some difficulties of segmenting cell cytoplasm from tissue samples.

One image of the cytoplasm was randomly chosen and manually segmented by three experts. In this example(Fig. 4.11), the experts were asked to segment the cell cytoplasm based on the image. They were able to see each channel of the image separately or merge them in any combination they wanted. Prior to the segmentation of the cytoplasm, they segmented cell nuclei from this image. The experts were given

	Expert A	Expert B	Expert C
Expert A	100	75.5	58.7
Expert B	75.5	100	63.9
Expert C	58.7	63.9	100

Table 4.2: The Dice similarity between three experts.

	Expert A	Expert B	Expert C	Iterative Tessellation	Tessellation
Expert A	100	92.3	92.8	86.8	84.4
Expert B	92.3	100	93.5	86.4	84.8
Expert C	92.8	93.5	100	86.2	84.5
Algorithm	86.8	86.4	86.2	100	92.1
Tessellation	84.4	84.8	84.5	92.1	100

Table 4.3: The Dice similarity between three experts and two computer algorithm.

instruction on how to use the software to draw it, but was not able to discuss their segmentation results with each other and they segmented the images based on their experience. They all had more than three years of experience working with this type of images. As shown in Fig. 4.11, the experts segmentations look different. The segmentation of the cytoplasm formed tessellation partitions the space between neighboring nuclei. In (d), the expert tends to draw a tighter line around nuclei and in (c) and (d) the experts tend to draw a larger cytoplasm around the nuclei. The Dice similarity between experts and the computer derived tessellation were compared with each other. The results are shown in Table 4.2. The similarity between experts range from 58.7% to 75.5%, which means that three experts find it hard to find the boundary of cells from this image.

However, in an image where experts can segment the cell cytoplasm, the results of the a single automatic watershed tessellation shows similar results to experts. A good quality image is given to three experts and they manually segmented the image. The image was also segmented using a watershed tessellation and a cytoplasm

segmentation algorithm, an iterative tessellation[2] The watershed tessellation showed similar results to the experts and the cytoplasm segmentation algorithm. For this reason, watershed tessellation was used for further analysis.

4.6 Quantitative feature extraction

We used a quantitative criterion to identify good and informative nuclei which required the area of nuclei to lie within a range from 1500 pixels to 8000 pixels. Features of cells were obtained from good and informative nuclei. Quantitative features such as total intensity of Ki67, the area of nuclei and cytoplasm were then computed as shown in Fig. 4.10.

4.7 Validation of the Algorithm

In order to validate the algorithm, three images were selected randomly from the imaging dataset. Each image contained more than 100 cells and they are manually segmented by three experts. The average of the manual results were computed and used as the ground truth. Nuclei and cytoplasm segmentation were validated by comparing the computed segmentation results with the ground truth. The validation results were analyzed by ROC analysis (Table. 4.4), Bland-Altman test and Hausdorff distance as shown in the following figures.

[2]This algorithm is originally designed by the author. Because it is not directly in use, the details can be found by contact the author directly.

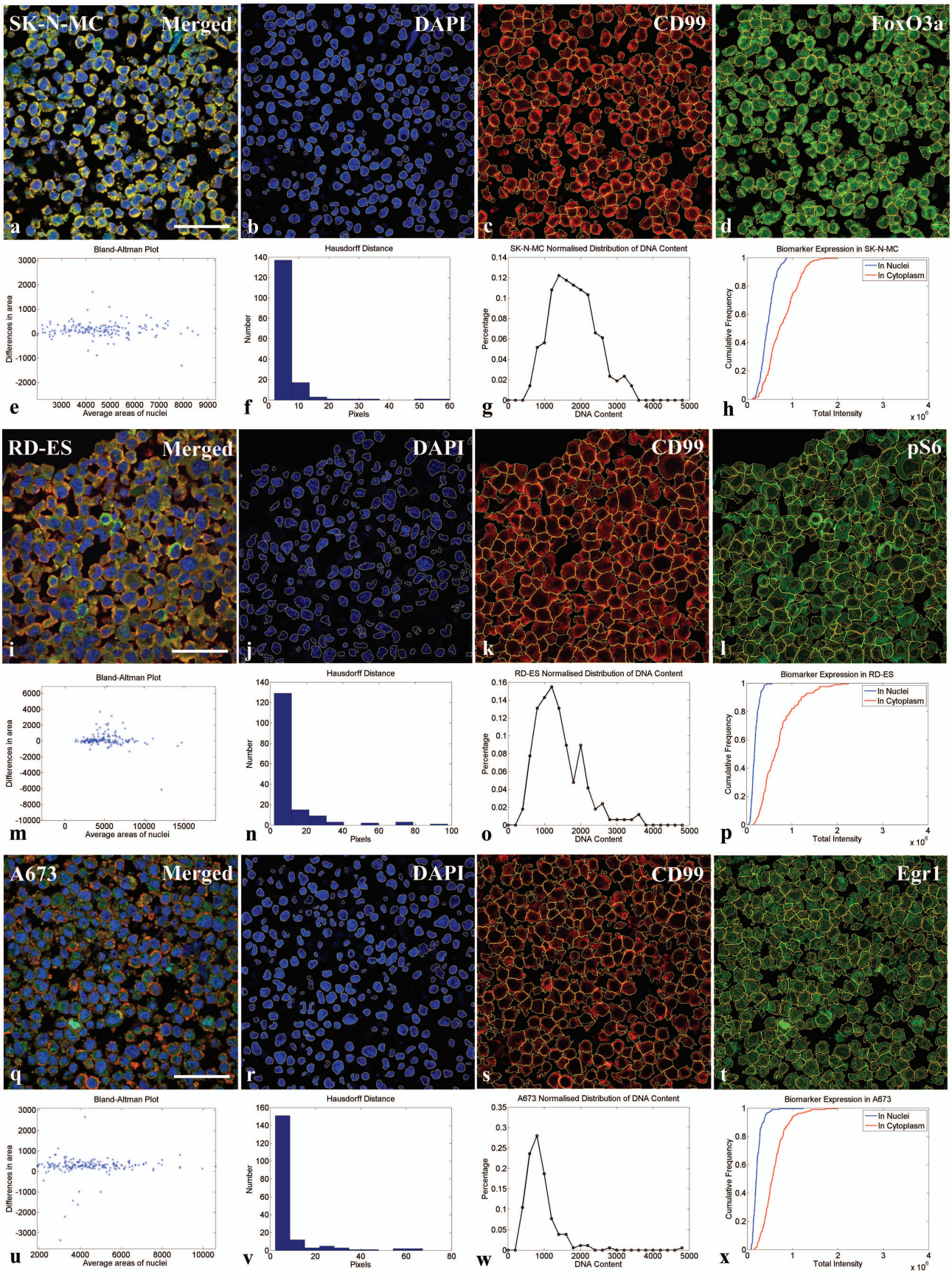

SK-N-MC
Merged
DAPI
CD99
FoxO3a
a
b
c
d
Bland-Altman Plot
Differences in area
Average areas of nuclei
e
Hausdorff Distance
Number
Pixels
f
SK-N-MC Normalised Distribution of DNA Content
Percentage
DNA Content
g
Biomarker Expression in SK-N-MC
In Nuclei
In Cytoplasm
Cumulative Frequency
Total Intensity
x 10^5
h
RD-ES
Merged
DAPI
CD99
pS6
i
j
k
l
Bland-Altman Plot
Differences in area
Average areas of nuclei
m
Hausdorff Distance
Number
Pixels
n
RD-ES Normalised Distribution of DNA Content
Percentage
DNA Content
o
Biomarker Expression in RD-ES
In Nuclei
In Cytoplasm
Cumulative Frequency
Total Intensity
x 10^5
p
A673
Merged
DAPI
CD99
Egr1
q
r
s
t
Bland-Altman Plot
Differences in area
Average areas of nuclei
u
Hausdorff Distance
Number
Pixels
v
A673 Normalised Distribution of DNA Content
Percentage
DNA Content
w
Biomarker Expression in A673
In Nuclei
In Cytoplasm
Cumulative Frequency
Total Intensity
x 10^5
x

Image	Expert A	Expert B	Expert C	Agreed	Computer	Accepted	Sensitivity	Specificity
1	212	231	228	221	182	177	83.9%	97.3%
2	213	237	219	202	212	190	94.1%	89.6%
3	175	189	206	175	168	161	92%	95.8%

Table 4.4: Validation of nuclei segmentation against three experts.

Image	Expert A	Expert B	Expert C	Agreed	Computer	Accepted	Edge	Accuracy
1	169	169	169	169	169	168	14	94.1%
2	201	204	200	200	204	204	8	98%
3	156	156	153	153	156	156	12	98.4%

Table 4.5: Validation of cytoplasm segmentation against three experts.

4.8 Conclusion

This chapter has successfully implemented an image processing algorithm to segment cell nuclei and cell cytoplasm. The difficult of the cytoplasm segmentation is highlighted in this chapter. An equal distance based watershed tessellation was found to be good for segmenting cell cytoplasm.

Validation of the segmentation algorithms showed good performance on the images in this study, which were also of relatively good quality. For lower quality images it would be important to control the quality of the tissue image before applying this quantitative analysis which is the subject considered in the next chapter.

Chapter 5

Classification of patterns in tissue samples with different quality

5.1 Introduction

One of the challenges in quantitative analysis of TMA comes from the difficulty of controlling the quality of a biopsy. Due to the variations in sample preparation and uncontrollable factors in experiments, the quality control of a tissue biopsy is an important step prior to quantitative analysis. Common types of poor quality images often contains stromal cells, DNA degradation and clumpy nuclei which can lead to errors in image segmentation and quantitative analysis.

Experienced pathologists typically assess the images visually to determine if they meet the standards necessary and can be subjected to computer analysis and data-mining. However, this procedure is time consuming and can introduce bias in sample selection. It is highly desirable to develop an automatic system to assess the quality of tissue sample images. However, traditional image analysis methods are not capable of modeling differences in tissue quality, while heterogeneity of samples exit in images. Due to heterogeneity of tissue, the morphology of cells varies a lot. It can be challenging to distinguish tissue quality related changes from a large amount of other changes.

In this chapter we analyse image regions of different quality using a computational method. The success of automatic classification is heavily dependent upon the choice of image features and this work studied the state-of-art image features and analysed their performance together with classification models. As a main contribution, this work discovered the features of each type of region and built an automatic system that can classify tissue with different quality.An early version of this work was presented at the International Workshop on Image-based System Biology, 2014.

5.2 The problem

In order to study tissue quality, a set of images were provided by a pathologist. A few examples are presented in the following figures with their characteristics discussed. The images are typically 2048×2048 pixels in size and immunolocalised with four fluorescent dyes (see Chapter 3, Section 1).

Previous research [23] on quantifying TMA images, has suggested that a good tissue sample needs to contain a certain amount of cells, relatively low background staining, clear nuclei boundaries, few stromal cells and a certain amount of tumour cells. A few examples are provided next to show the complexity of the tissue quality problem. The first three examples are excluded from quantitative analysis by pathologists because there are regions in the image with poor quality. The definition of poor quality is explained in detail with the examples next.

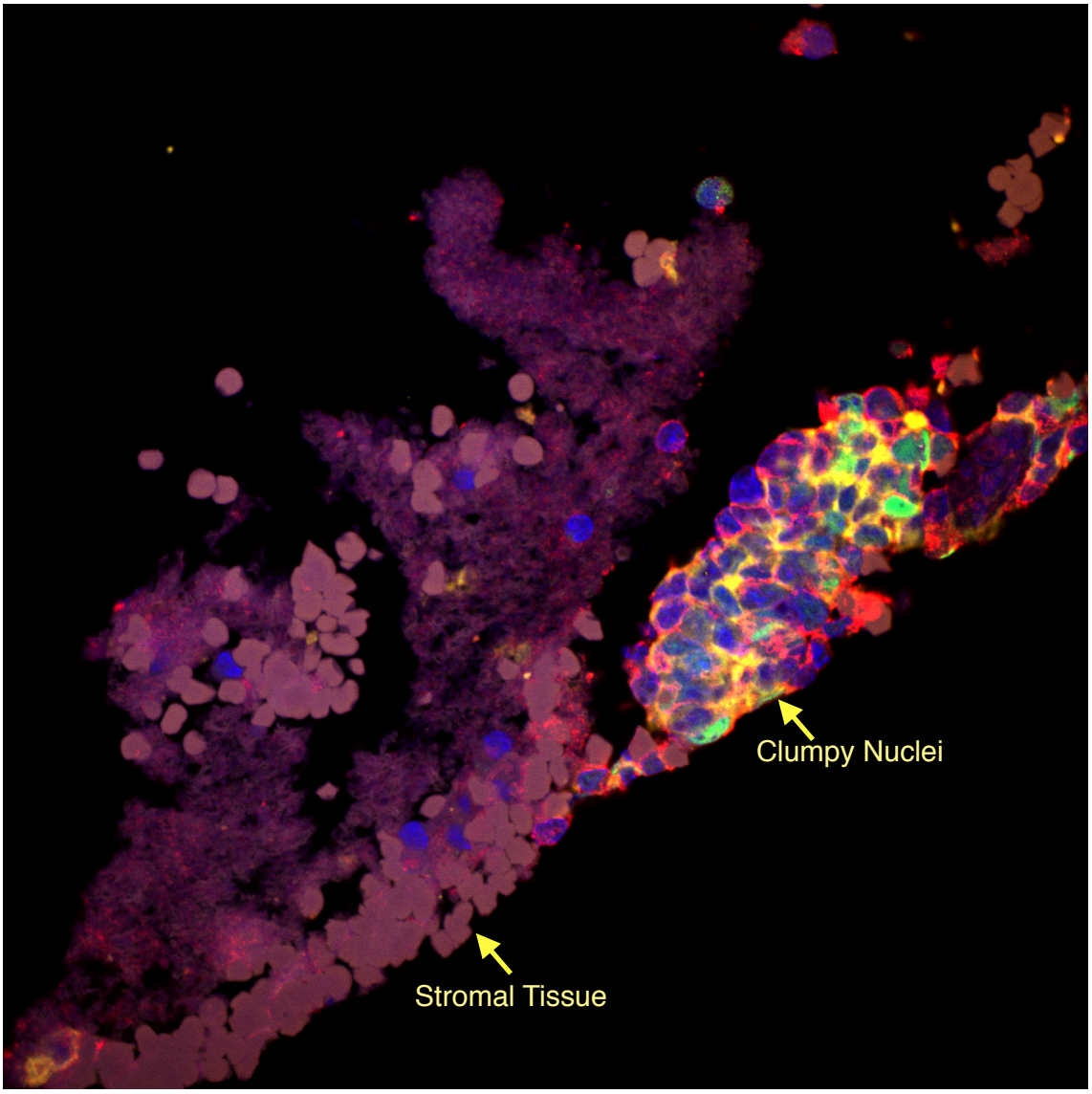


Figure 5.1: An example of a poor quality TMA. There are a large amount of stormal cells at the left side of the image and background staining. The right bottom corner and left upper corner have no cells and there are not many tumour cells in this image.

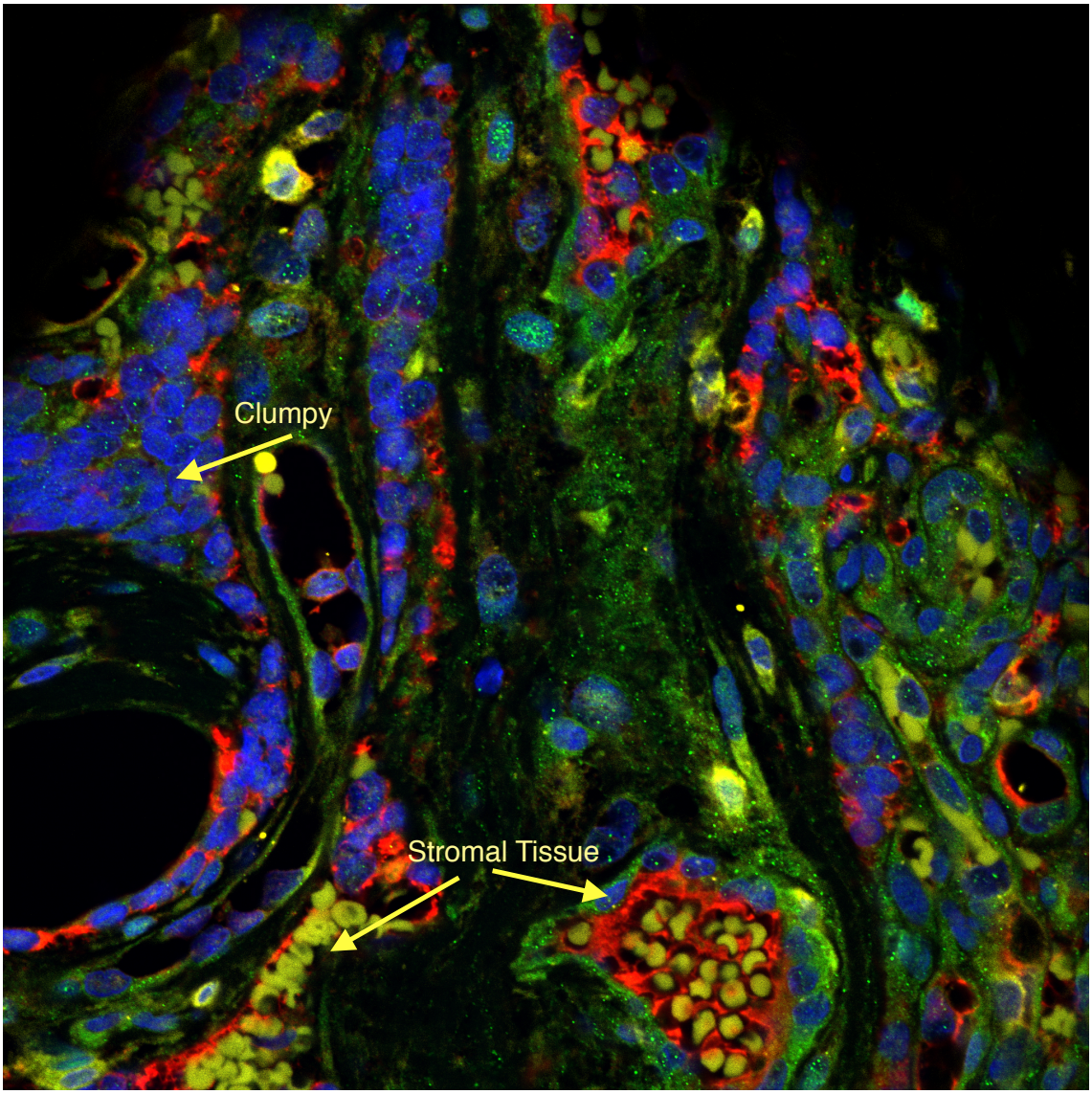


Figure 5.2: An example of a poor quality TMA. There are stromal cells in different regions and the cells form big clumps.

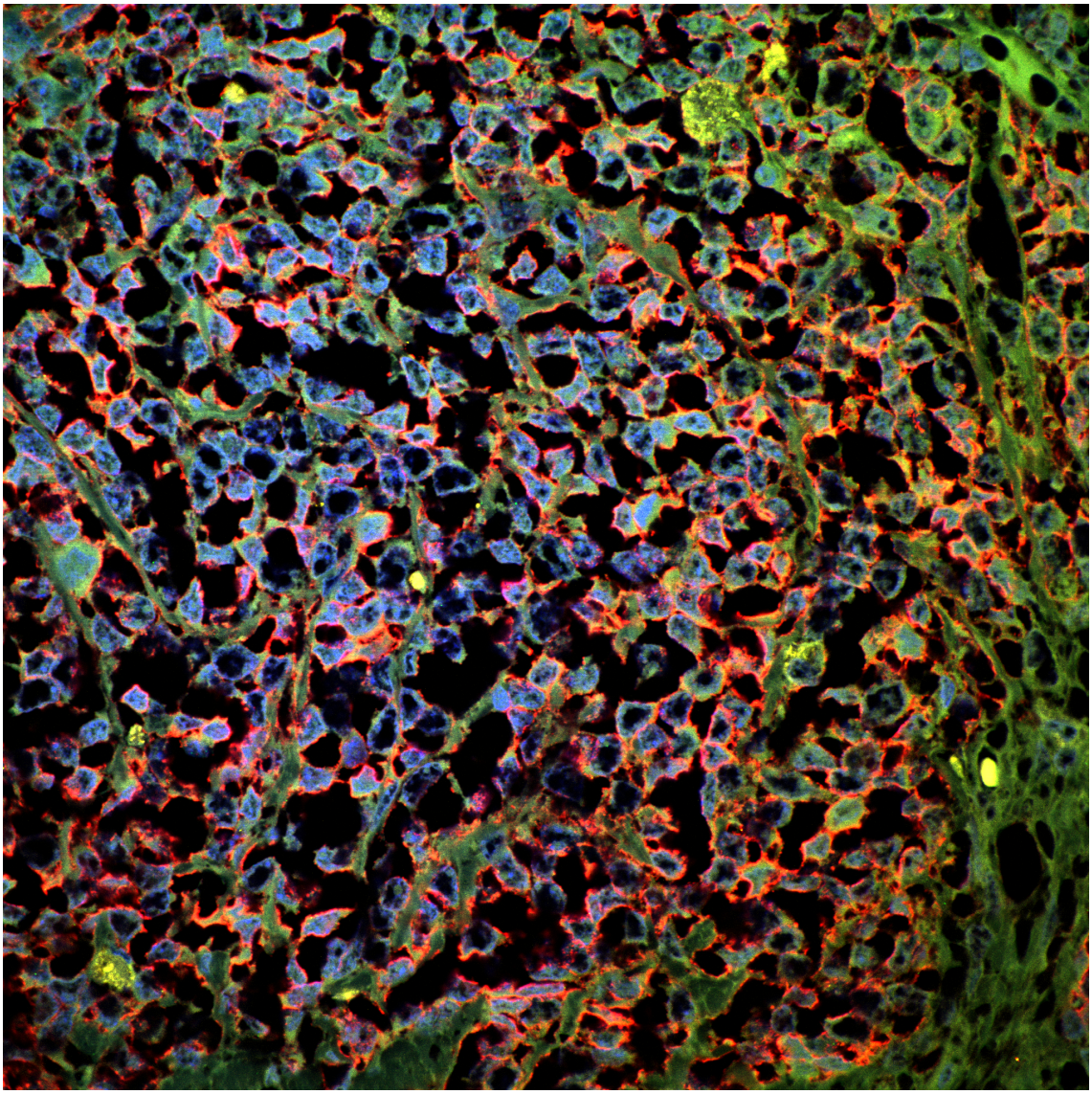


Figure 5.3: An example of a poor quality TMA. The cell nuclei has less DAPI staining compared with the cell cytoplasms. It might be a result of degraded nuclei.

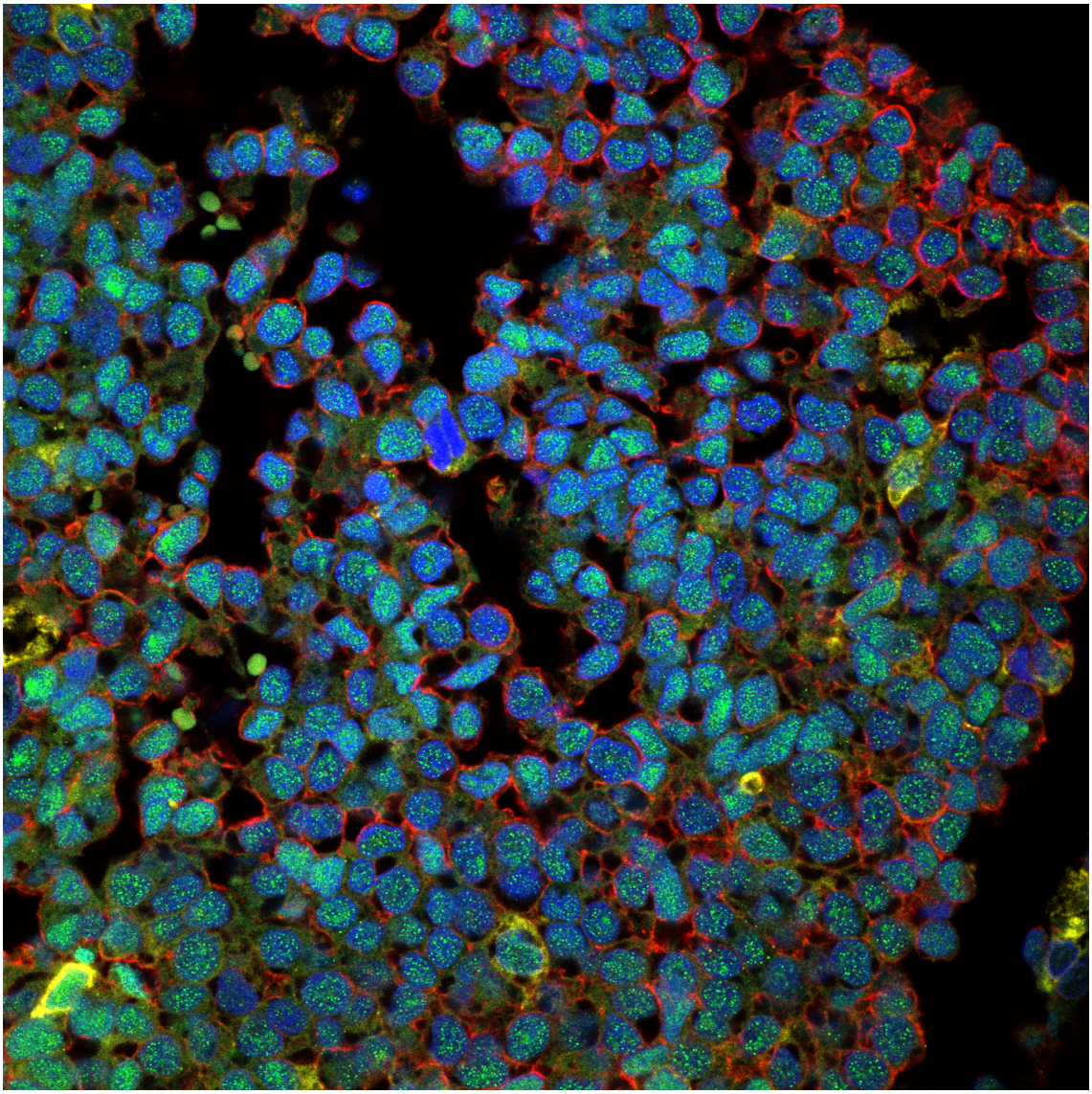


Figure 5.4: An example of a tissue image with good quality. There are many tumour cells and few stromal cells in the image and the boundaries between nuclei are clear.

In order to show the regions in the image clearly, a few patches are taken from the images and presented in Fig. 5.5. In Fig. 5.5(a), an example of cells with clear nuclei is presented, we named this group as 'normal' nuclei. However, in Fig. 5.5(b) the DNA content of nuclei seems to come out of the nuclei, which might be a result of DNA degradation. This case is named as 'degraded' nuclei. In the example shown in Fig. 5.5(c), the nuclei of cells form clumps and the boundaries between nuclei is hardly visible. It is not possible to separate individual nuclei from the clump. These cells are named as 'clumpy' nuclei. In Fig. 5.5(d), the cells in this image are not tumour cells. In a tissue sample, there might be stromal cells and they are named as 'stromal' tissue. In tissue images, these different type of cells exist in one image, which might be located at different regions and there is no clear boundary between each regions. If the region has a large amount of stormal cells, degraded DNA, and very clumpy nuclei, they are considered poor quality. If the sample contains a large amount of tumour cells, it is considered as good quality.

As shown in the example images, even though the difference in quality can be observed by pathologist, it is still challenging for an automatic algorithm to recognise these differences. Ideally it could be useful to discover a way to quantify these differences computationally. Because the cells and cell clusters form different patterns, it is interesting to know if these textures are related to tissue quality. The main aim of the work described in this chapter was to study if there exists image features that can differentiate regions with different quality. In addition, a classifier was built to classify regions with different quality.

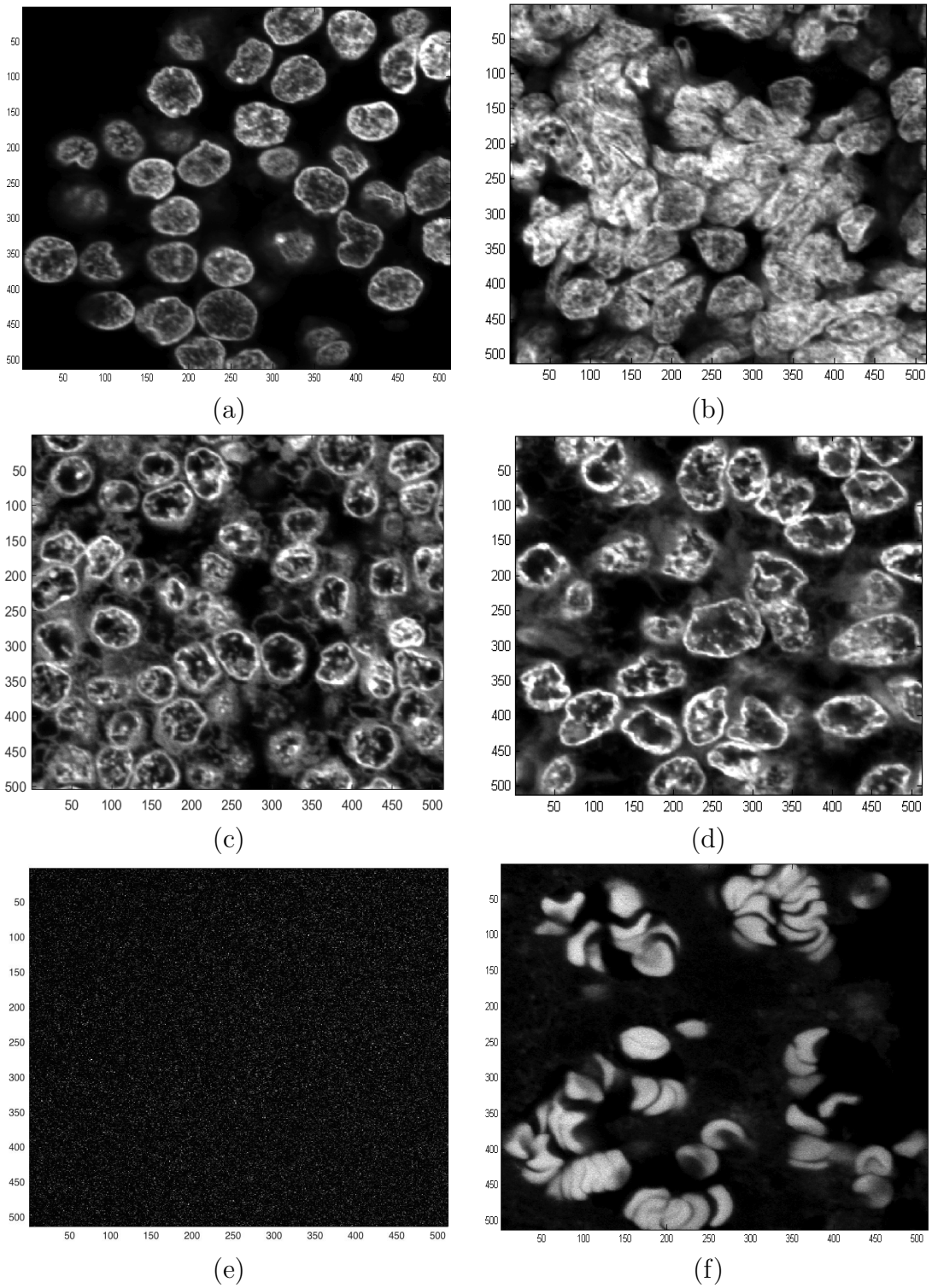


Figure 5.5: Examples of image patches taken from original tissue images. (a) An example of 'normal' nuclei. (b) Nuclei are clumpy and overlapping. (c) The DNA in the nuclei sample is degraded. (d) The nuclei in this image showed signs of degradation, but seems less severe compared with (c). The size and shape of nuclei showed heterogeneity with them in (c).(e) Background staining and noise. (f) Stromal tissue in the sample is stained by the fluorescent probes.

5.3 Methodology

In order to study image features related to different type of tissue, the pathologists manually selected the image regions from the original image and labelled them. The selection of image regions was performed as explained in detail in the following text. Image features and classification models were built to recognised these regions. The previous literature [46, 109] suggested texture features such as Gabor filter, co-occurance features and first-order statistics have discriminative characteristic. However, the state-of-art descriptors for cell images such as scattering convolutional network (SCN), maximum stable extremal regions (MSERs) etc,have not been studied before. This work studied their discriminative character in this application.

5.3.1 Data Annotation

One of the challenges in this work is to define a reasonable way of labelling the different regions in the imaging data. It is possible for pathologists to identify these regions, However, in practice, it is very hard for them to draw a precise boundary around different regions. In order to enable a pathologist to label the data, a user interface was designed which allows the pathologist to manually select and save the regions of different quality. In this work, pathologists placed square bounding boxes around the regions they identified.

The size of the image regions identified by pathologists are highly variable. The assumption is that there is a certain window size above which we are able to visually distinguish between regions with different quality. However, if the window is too large, there are multiple objects in the window and we will no longer be able to assign a single label to the window. This provides us with the upper limit of window size. In this study, multiple scales are used.

Five different classes of image patches were identified from the data by a pathologist and manually grouped into five groups. The five classes are clumpy nuclei (e.g. figure 5.5(b)), stromal tissue(e.g. figure 5.5(f)), degraded DNA (e.g.figure 5.5(c)), normal nuclei (e.g.figure 5.5(a)) and background (e.g.figure 5.5(e)). The datasets were manually selected by two experts (a pathologist and an engineer with 3 years of experience in quantitative biological imaging).

5.3.2 Image Features

Image features are key to the success of classification models and the discriminative features need to be discovered. The author has tried to included image features that are most relevant to the application, and they are computed and listed in Table. 5.1. Each of these features is explained in detail in the following text. These features were put together in the order shown in Table. 5.1, to form a single feature vector for further analysis.

Fractal Dimension [133] This measures the complexity of a texture and comparing changes of pattern with its scale. The calculation of fractal dimension used differential box-counting algorithm and the mean, standard deviation and lancunarity of fractal dimension are computed.

Scattering Convolutional network (SCN) [22] This is a class of convolutional network, which filters the image with a family of wavelet transforms and is computed with a filter bank of dilated and rotated wavelets. In this work, the Morlet wavelet (or Gabor wavelet) is used [22]. A Morlet wavelet is a complex exponential multiplied by a Gaussian window. The equation is:

$$\Psi(u) = C_1(e^{i\xi u} - C_2)e^{-|u|^2/(2\sigma^2)} \tag{5.1}$$

Here u is the pixel location, and C_1, C_2, ξ, σ represent constants that define a particular filter. Specifically, C_2 is adjusted so that $\int \Psi(t)dt = 0$. ξ controls the rotation of the filter, while σ controls the bandwidth of the band-pass filter. By tuning the rotation and bandwidth of the wavelet, a particular filter λ among this family of filters is obtained. The energy of an image I convolved with Ψ_λ is given in equation 5.2,

$$U = |I * \Psi(u)_\lambda| \tag{5.2}$$

The image is convolved with a network of filters and a particular path p of it is computed (equation 5.3).

$$U[p] = |||I * \Psi_{\lambda 1}| * \Psi_{\lambda 2}| \cdots * \Psi_{\lambda m}| \tag{5.3}$$

The energy within each path is obtained by $U[p]$ and each path produces a representation of the image. By computing many convolutional paths, an image can be represented by the energy of each path.

Auto-correlation This measures the sharpness of the image. The auto-correlation matrix of the image is calculated as an indication of the focus of the object [11]. For an image I, the Fourier transform of the image is $F(u, v)$ and F^* is complex conjugate of F. The equation for the autocorrelation matrix of image A is (in the Fourier domain)

$$A(u, v) = F(u, v)F^*(u, v) \tag{5.4}$$

The two measurements, slope and volume under the peak, are taken from the auto-correlaton matrix to measure the shape of the auto-correlation.

Intensity distribution The intensity histogram of image patches is computed and several parameters are derived from it; the mean, variance, skewness and kurtosis are measured.

Major edge density (MED) In order to assess the level of detail in a cell nuclei, the spatial domain analysis of image features is taken into consideration [44]. The density of major edges in an image region is computed, starting with a Canny edge detector [77] which compute the edges of image. The major edge density is the number of major edges found by the Canny edge detector divided by the total number of pixels in the region (Equation 5.5).

$$\delta = \frac{\sum_{pixel \in edge} I}{M \times N} \tag{5.5}$$

M and N are the width and length of image I. The edge density is proportional to number of cells and the texture of cells in the image.

Entropy and entropy upon predefined direction The Shannon entropy of a image is computed. $P(i)$ is the probability density function of intensity value i and the Shannon entropy[1] H is given as,

$$H = -\sum_{i \in I} P(i) \times log_2(P(i)) \tag{5.6}$$

In order to compute entropy along predefined directions, the Rényi entropy [57] upon discrete space-frequency distribution $P(n, k)$ can be computed as,

$$R_3(n) = -\frac{1}{2} \log_2(\sum_n \sum_k P(n, k)) \tag{5.7}$$

Here $P(n, k)$ is the probability density function of a position n, and frequency variable k and entropy is calculated by:

$$R_3(\theta_s) = \sum_n R_3(n, \theta_s)/A; \tag{5.8}$$

[1]Maximum is $log_2(N)$, N is 256 in our images.

Here, A represents the image size, $\theta_s \in [\theta_1, \theta_2, ..., \theta_s]$ represent s different orientations taken to measure entropy. Shannon entropy and the variance of along different direction are used as image features.

Correlation : this measures the correlation coefficient between pixels in the DAPI channel and in the CD99 channel. This feature is needed because sometime nuclei of stromal cells can be observed in the DAPI channel and the CD99 channel, while normal nuclei can only be observed in DAPI channel. The correlation coefficient is computed using:

$$\rho_{x,y} = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2 \sum_{i=1}^{n}(y_i - \overline{y})^2}} \tag{5.9}$$

x_i is intensity value of pixel i in DAPI channel and y_i is the intensity value of pixel i in CD99 channel.

Maximum stable extremal regions (MSER) [105] : here we compute the regions in the image which remain connected over a set of thresholds. The algorithm that we implemented, uses three levels of level set to compute the regions associated with each level and the differences in area and shape between each level is computed. These differences are constructed into a feature vector.

Histogram of oriented gradients (HoG) [39] : computes the distribution of directions of intensity gradients. The image is divided into small cells and within each cell, the histogram of gradient directions is computed. The maximum and mean of this histogram are obtained and formed into two vectors. These vectors vary with translational changes of the object.

5.3.3 Classification models

Multi-class Random Forest [20] Random Forest is an ensemble of multiple decision trees. A random forest model is built of several decision trees. Each decision tree

starts with a random subset of the features. At a parent node, the model examines all possible binary splits on every predictor and then selects a split that optimises the impurity of nodes (classification). Then the node is split into two child nodes and the process recursively repeat on the child nodes.The algorithm stops when a node is pure, the maximum depth of the tree is reached or other stopping rules. The voting of all the trees are then combined together.

Multi-class Support Vector Machine (SVM) [38] In the feature space of the datapoint, the support vector machine constructs a hyperplane or hyperplanes to classify these data. A support vector machine can be linear or non-linear depending upon the choice of kernel function. A multi-class SVM can be constructed using a one-against-all method [72]. A linear SVM is used in this study to explore the image features.

5.4 Experiment

The pathologist manually labeled the regions with different qualities and discovered empirically that most of the patches could be labelled with two sizes of bounding box, 512 × 512 pixels (named level 1) and 128 × 128 pixels (named level 2)[2]. The clumpy nuclei tended to form larger regions, while most of the stromal cells can be observed using smaller window. The two scales of windows were used to label the image region and extract image features from these two different scales. The training dataset contained image patches with two sizes. For the level 1 (patch size 512 × 512 pixels), 60 patches of each group were included in the training dataset, giving a total of 300 image patches. For the level 2 (patch size 128 × 128 pixels), 120 patches in each group were used giving in total 600 image patches. For both levels, the discriminative image features were computed and selected and classification models were built. The

[2]A typical cell nucleus is about 40 × 40 pixels. A level 2 image patch on average may contain 4 nuclei.

Features	Vector (Level 1)	Length (Level 1)	Vector (Level 2)	Length (Level 2)
Fractal Dimension	1-3	3	1-3	3
SCN	4-802	799	4-394	391
Auto-correlation	803-805	3	395-397	3
Intensity	806-809	4	398-401	4
Edge Density	810	1	402	1
Entropy	811-814	4	403-406	4
Correlation	815	1	407	1
MSER	816-821	6	408-413	6
HoG	822-823	2	414-415	2

Table 5.1: The break down of the complete descriptor.

complete breakdown of the image features are listed in Table. 5.1. Less SCN features were used because the level 2 image patches are smaller level 1 patches.

5.4.1 Feature Selection

The image features computed are listed in Table. 5.1. These features were put together according to the order shown in Table. 5.1, to form a single vector for further analysis. It is essential to select features that are discriminative and to exclude noisy features.

The features were selected using a sequential feature selection (also called wrapper [17, 85, 63]) method. The sequential feature selection method starts with an empty set of features and then sequentially adds features until there is no improvement in prediction. 10-fold cross-validation was used to assess the performance of prediction. Because this feature selection method is optimized with the classification models, two classification models were compared.

The results after applying support vector machine model and random forest models are shown in Figure 5.6. The blue line with the ring shows the improvement of

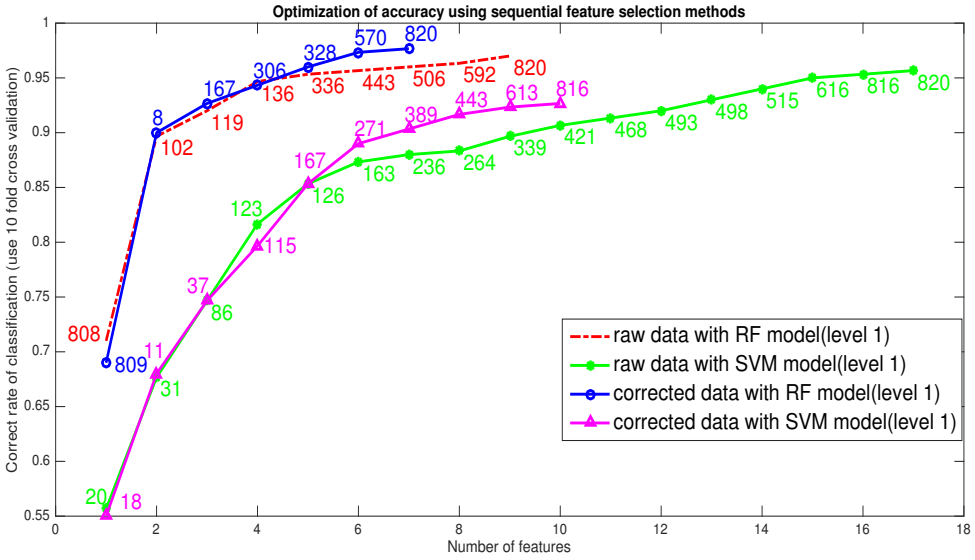
prediction by sequentially adding in more features. The details of each feature is explained in detail in the section 5.4.2. The magenta line shows the sequential feature selection using a support vector machine. The performance of the support vector machine in this experiment is not as good as a random forest model. The features that were selected by the support vector machine were found to be different than for the random forest. The number of features used by the support vector machine was larger than required by a random forest (RF) model of equivalent accuracy.

The feature importance[3]obtained from the random forest model is also shown in Figure 5.6. This importance is represented by an arbitrarily scale generated by the RF; the higher this number is, the more important it is for RF to classify the data. This confirms the observation of previous studies [146, 58, 8] that the importance values of the most discriminant variables are not necessarily higher than a less discriminant one, when these predictors may correlate with each other.

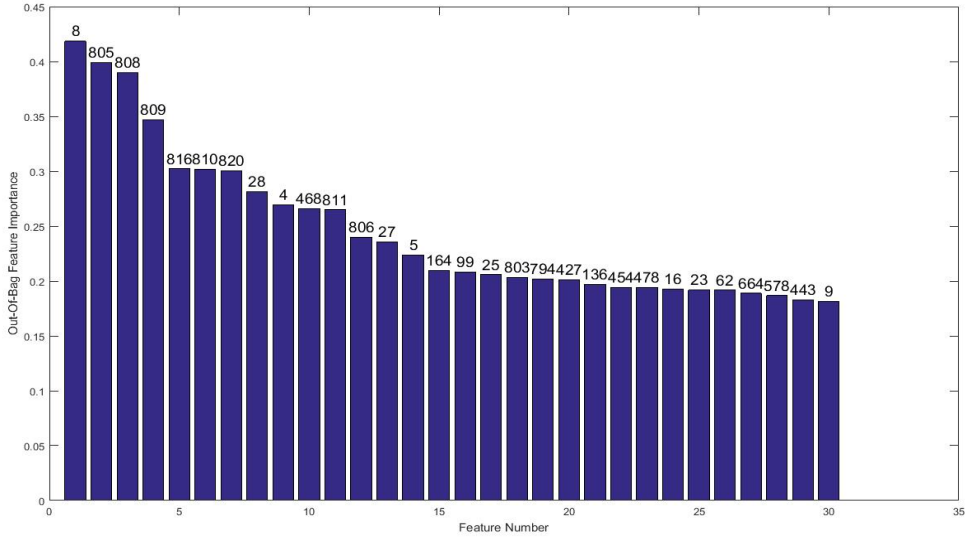
However, the features selected by SVM and by RF models are not same. This might be because SVM in this study used linear kernel and Random Forests is not a linear model. This also suggests that the high dimensional image features are highly complex. In order to build a practical model, it is important to use discriminative features to build the classifier.

In addition, this experiment also tested the impact that uneven illumination had on recognition of tissue images. The red line and green line show the selection of features without illumination correction of these regions. As explained in Chapter 3, uneven illumination across images introduces dark corners and leads to errors in quantitative analysis. The corrected data and the original data used different features

[3]Details on the computation of feature importance can be found in Breiman's paper [20].

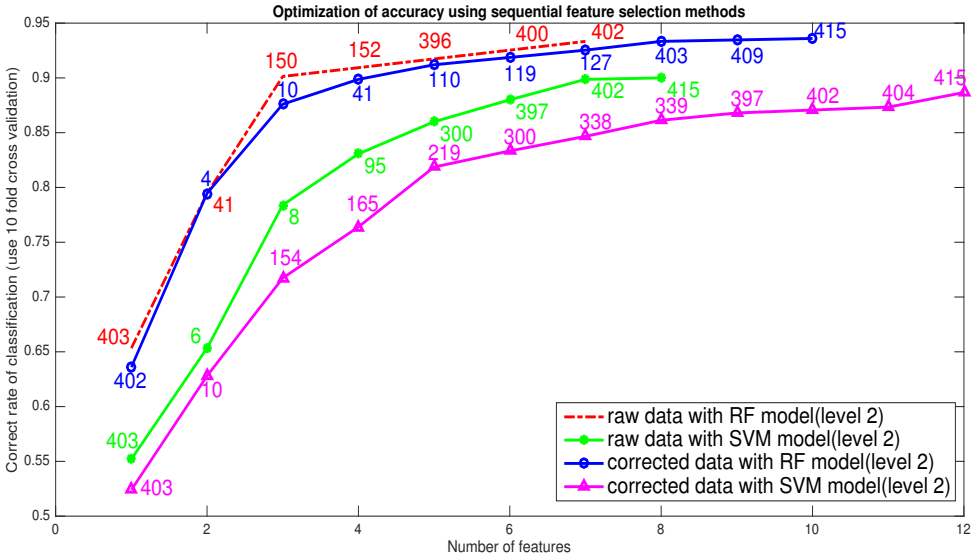


(a) Sequential Feature Selection.

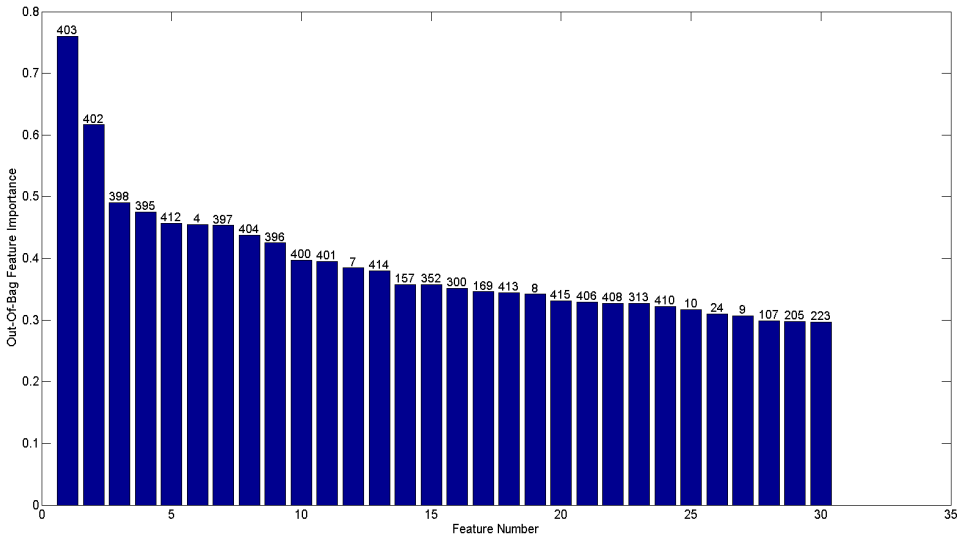


(b) Feature importance by RF

Figure 5.6: The sequential feature selection and feature importance of all the variables for image patch Level 1. (a) By adding the features sequentially, the model improved classification accuracy. (b) Feature importance obtained by a random forest model. The number corresponds to the location of the feature in the feature vector in Table. 5.1.



(a) Sequential Feature Selection.



(b) Feature importance by RF

Figure 5.7: The sequential feature selection and feature importance of all the variables for image patch Level 2. (a) By adding the features sequentially, the model improved classification accuracy. (b) Feature importance obtained by a random forest model. The number corresponds to the location of the feature in the feature vector in Table. 5.1.

for prediction. In both RF and SVM model for level 1, the number of features needed to classify the original data are larger than for the corrected images, which might indicate that the original data is more complex. The prediction on corrected data is very similar to that of the raw data, suggesting that the correction of illumination preserves key information in image.

However, in level 2 models (Fig. 5.7), the features of raw data showed similar results in RF models and much better performance in SVM. Level 2 image patches are much smaller and contains local information about cell nuclei. When the level 2 image patches are taken from the original image, and grouped into five classes, the global shading effect on their original image did not affect which group a image patches should be labelled. By removing the illumination effect, more features are needed to build a classifier. This result indicates that for smaller patches, there is less effect of shading and here correction is not so critical.

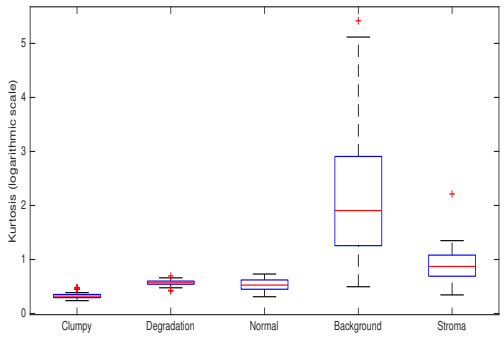
5.4.2 Insight into Image Features

Random forest Level 1 model. The most important features for the classification are SCN, MSER shape and intensity kurtosis. In the random forest model, the first feature that picked up by the algorithm is intensity kurtosis and its value across data point are shown in Fig. 5.8 (a). The 'background' group has a higher kurtosis and a larger range of values. This may suggest that background contains less useful information. The second (shown Fig. 5.8 (b)-(e)) to the sixth feature found by the algorithm is the energy response of the SCN. These features improve the prediction accuracy from 70% to around 90%. The SCN features are descriptors of tissue texture. The tissue samples have different textures related to their quality. The last feature is mean area of the MSER (shown Fig. 5.8 (g)). The stromal tissue has a

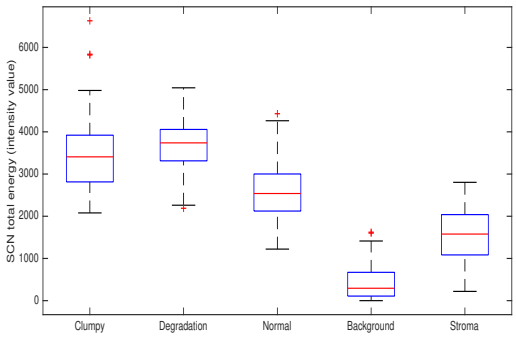
large mean MSER area because they can form clusters.

Random forest Level 2 model The most important features for classification are major edge density, SCN, Shannon entropy, MSER shape and mean of HoG (Fig. 5.9, Fig. 5.10). The major edge density seems higher in clumpy and degradation group while lower in normal and stroma groups. However, when it comes to background, the major edge density has a wide range and was found to not be a useful discriminator. While the Shannon entropy shows that the background has a lower amount of information in it. The SCN features was good for describing texture of tissue nuclei.

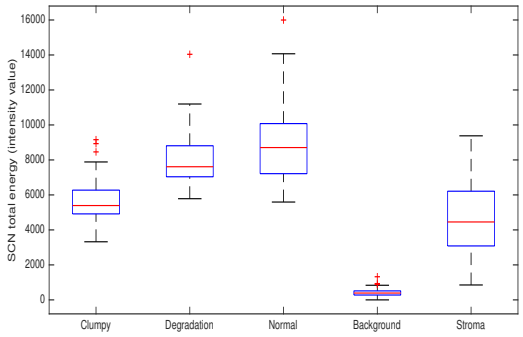
SCN features In both level 1 and level 2 models, the SCN features were found to be important. By iteratively, convolving the image patches with several filters, the changes of textures of cell nuclei in image patches can be described by these features. However, some of the variations of textures are due to heterogeneity of samples (Fig. 5.5(c) and (d)). In this experiment, it is important to use SCN features together with other types of non-texture features, not texture related features, such as intensity, entropy etc, to built a classifier.



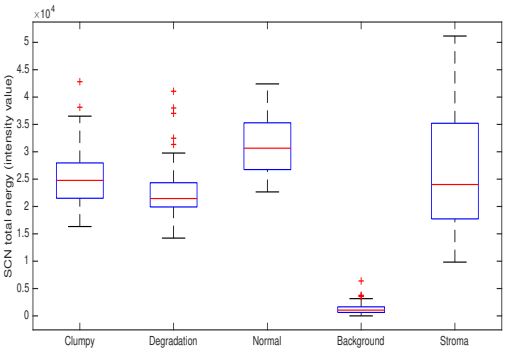
(a) Kurtosis of the intensity distribution



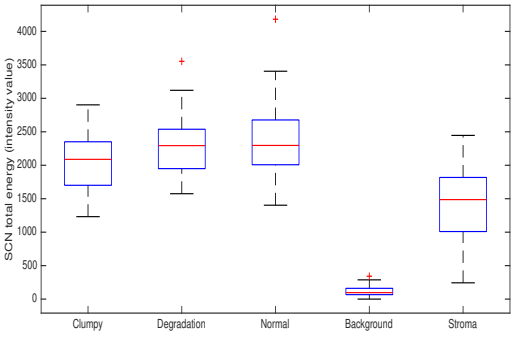
(b) SCN feature 8



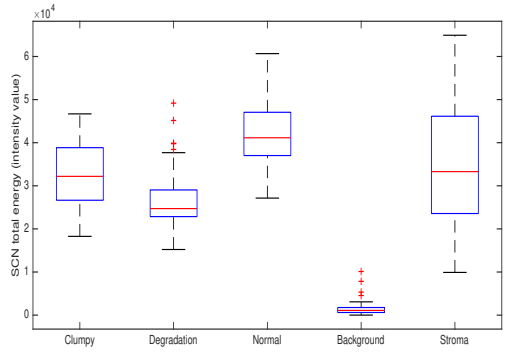
(c) SCN feature 167



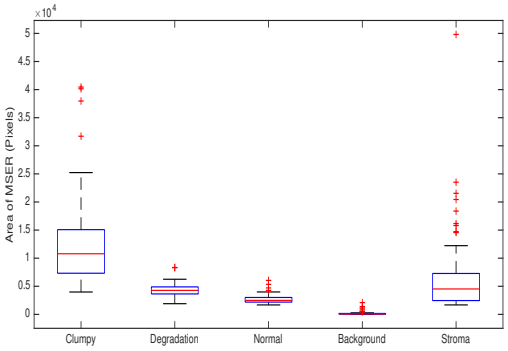
(d) SCN feature 306



(e) SCN feature 328

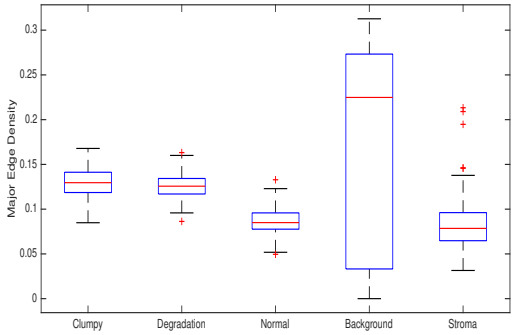


(f) SCN feature 579

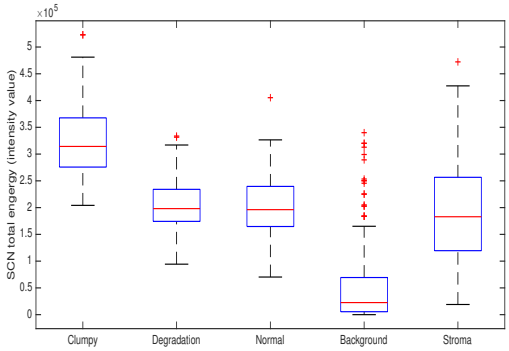


(g) MSER average area

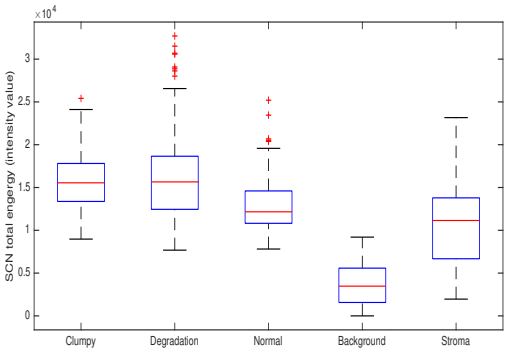
Figure 5.8: The box plot of discriminative features for level 1 patches.



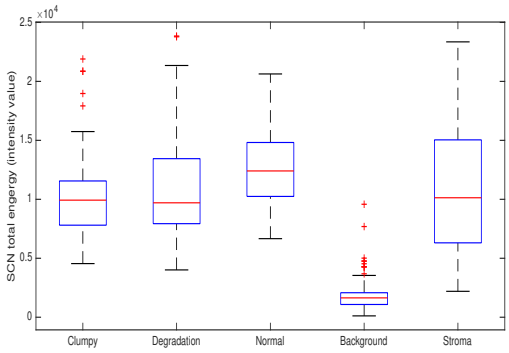
(a) Major Edge Density



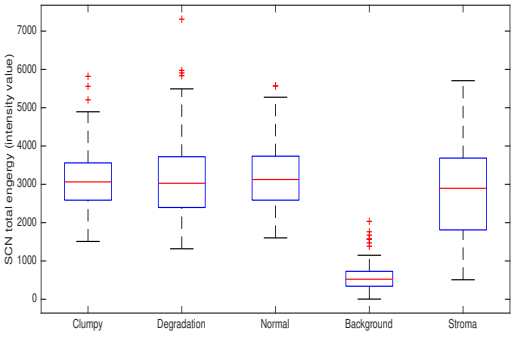
(b) SCN feature 4



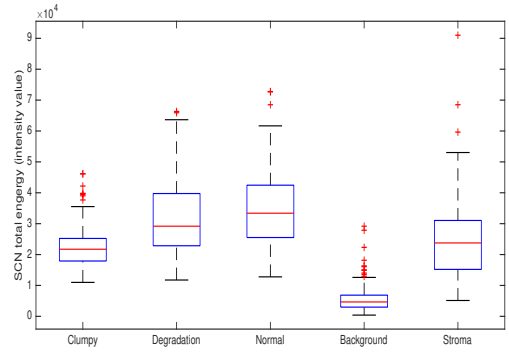
(c) SCN feature 10



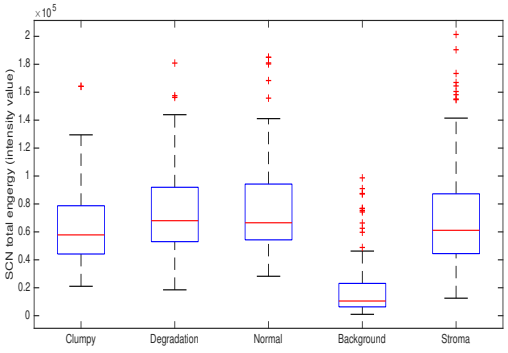
(d) SCN feature 41



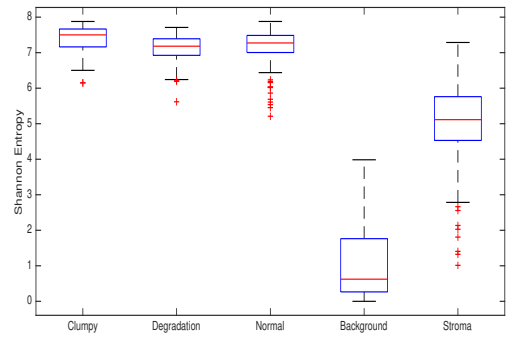
(e) SCN feature 110



(f) SCN feature 119

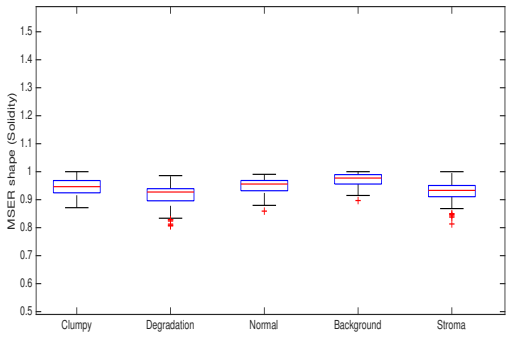


(e) SCN feature 127

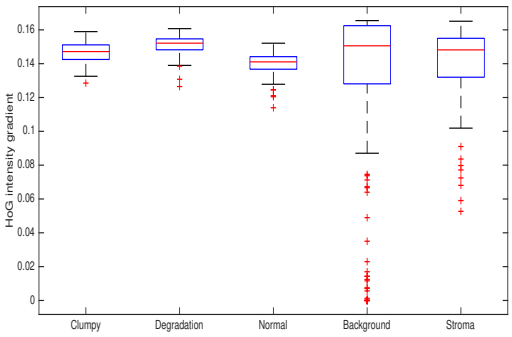


(f) Shannon Entropy

Figure 5.9: The box plot of discriminative features for level 2 patches.



(e) MSER Shape



(f) HoG Mean

Figure 5.10: The box plot of discriminative features for level 2 patches (continued).

Model	No.Trees	Left out%	Accuracy (mean, std%)
RF(Level 1)	100	75	91±2.4
RF(Level 1)	100	50	94±2.2
RF(Level 1)	100	25	95±2.5
RF(Level 2)	150	75	88±1.3
RF(Level 2)	150	50	90±1.4
RF(Level 2)	150	25	91±1.8

Table 5.2: Cross validation of Random Forest models

5.4.3 Random Forest Classification

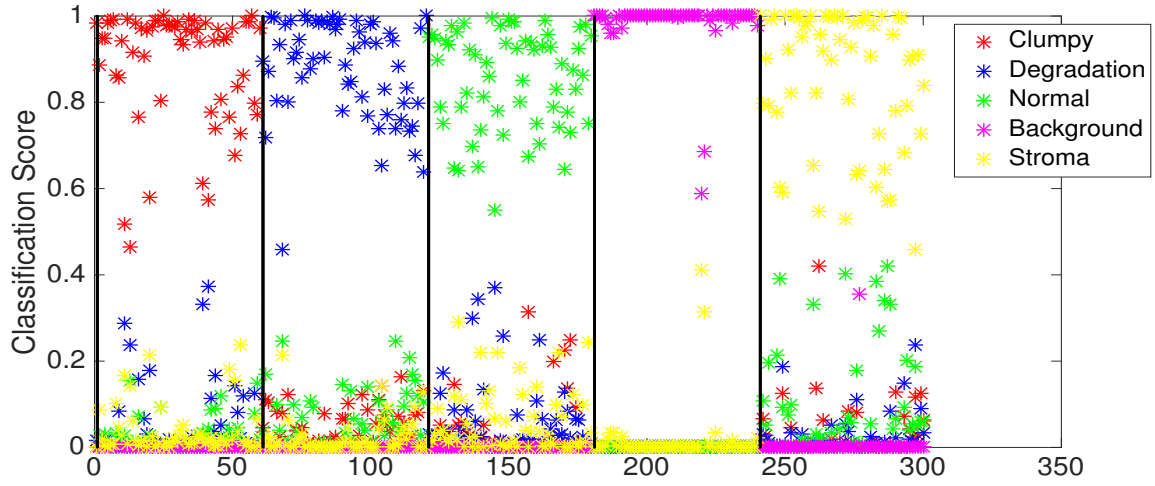
In order to build a classification model, a random forest classification framework is adopted and trained to classify the images using the classes of features selected. The split criteria used at each node was Gini's Diversity Index,

$$G = 1 - \sum_{i} P^2(i) \tag{5.10}$$

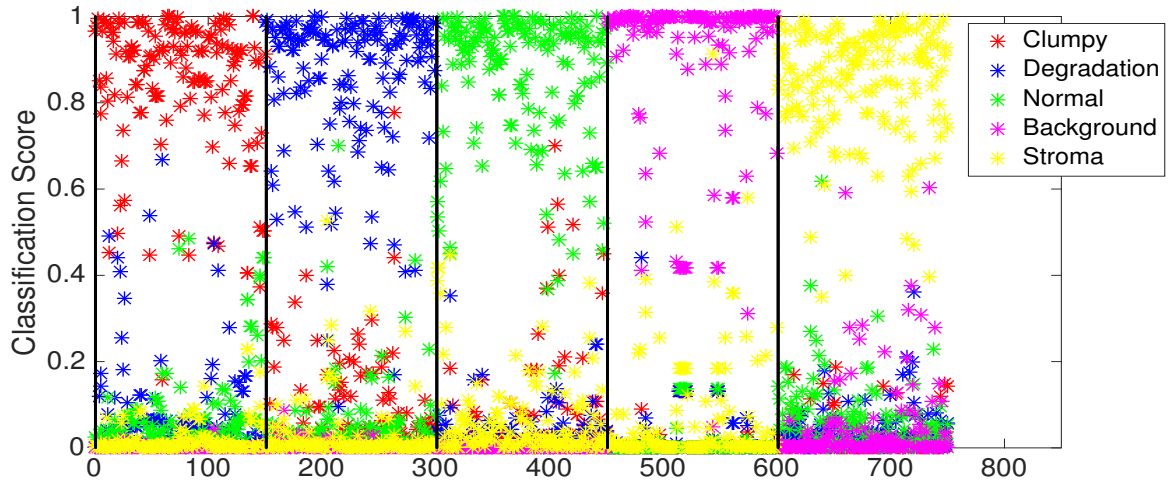
for classes i at the node, summed over the probabilities $P(i)$ that i will reach the leaf.

In order to avoid over-fitting, the number of trees and the depth of trees were optimised and repeated leave-one-out cross-validation (randomly selecting 25% to 75% of the data hold out for testing, repeated 100 times) was used while building Random Forest models. The results of different models are summarised in Table.5.2.

Observe that the level 1 random forest classification had a relatively higher accuracy, however, the size of the patches are too large, which is not very helpful when smaller regions need to be identified. The level 1 results are better than those of level 2, because the variances in image patches increases as the image patches become smaller.

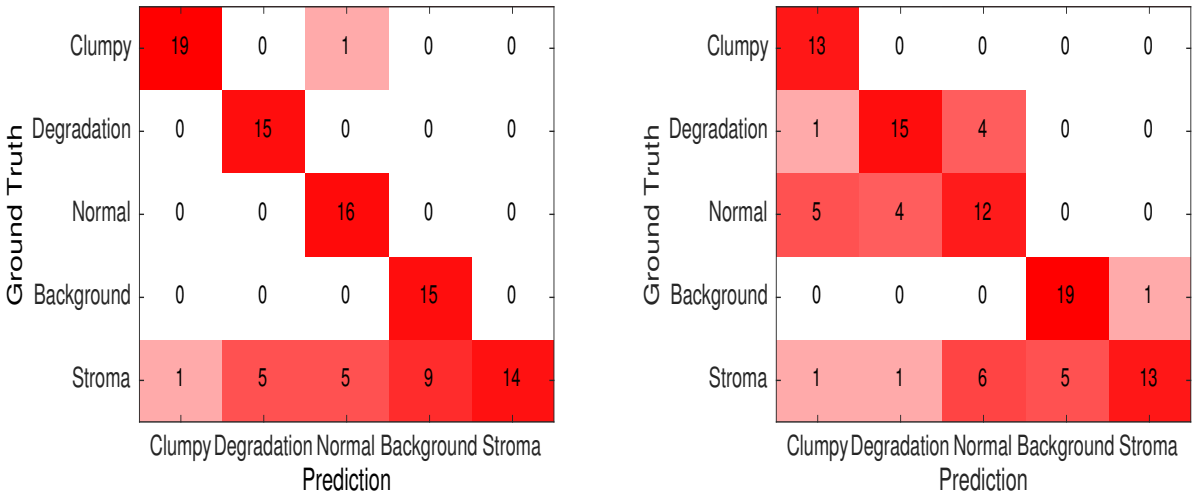


(a) Level 1



(b) Level 2

Figure 5.11: The classification scores for level 1 and level 2 image patches. There are 5 sets of 60 data points (level 1) or 150 data points (level 2) sequentially ordered as shown in the (a) and (b). The different colour represent the probability that the data points belong to a particular class. The red colour represents clumpy nuclei class, the yellow colour represents the stromal tissue class, the blue colour represents the degraded nuclei class, the green colour represents the normal nuclei class and the magenta colour represents the background class.



(a) RF Level 1 (b) RF Level 2

Figure 5.12: Classification results on 100 new images patches.

5.4.4 Validation

The classifier was tested on 100 independent samples and image patches were manually selected by experts. The RF (level 1) and RF (level 2) were tested. The RF(level 1) achieved 79% of accuracy, while RF (level 2) achieved 72% of accuracy. Figure 5.12 shows the performance of the classifiers as confusion matrixes. In the level 1 model, 19 'clumpy' image patches were classified as 'clumpy', while 1 was classified as 'normal'. 15 'degradation' examples, 16 'normal' examples and 15 'background' examples were classified correctly. However, 1 'stroma' is classified as 'clumpy', 5 'stroma' is classified as 'degradation', 5 as 'normal' and 9 as 'background'. Among all the labelled groups, level 1 classifier have difficulty with 'stroma' group. However, the overall performance was good. The level 2 classifier had really good performance on 'clumpy' and 'background' groups, but found classifying 'normal', 'degradation' and 'stroma' groups more difficult.

For both level 1 and level 2 classifier, the 'stroma' group was very difficult to classify. This is because the 'stroma' cells could also form clusters and show some features of 'clumpy'. In addition, stromal cells could also have low intensity values that make

it hard to classify. In addition, the stormal tissue class might contain blood cells and lymphocytes, as these cells have a smaller size and they are likely to be mixed with cancer cell samples. Therefore, stromal tissue images often have mixed cell types.

5.5 Discussion

Image regions with different quality used in this study were manually selected. This work has shown the ability to describe the differences of tissue quality of selected regions. However, in order to build a fully automatic method, it is important to automatically find the region of interest.

The two different levels patch sizes investigated in this study, have shown that different type of features are discriminative at each level. However implementation of feature extraction algorithm are normally time consuming, due to this limitation, only limited amount of features can be tested in this experiment. In this experiment, image features were mainly extracted from nuclei channel. Image features extracted from other channels and multiple channels might improve the classification results. Using more or different image patch sizes might improve the performance of classification models.

Increase the number of image patches might also help to improve the selected subset of features [76]. The number of image patches in this study was limited due to manually labelling the data being time consuming. The features selected by this sequential feature selection method have shown good discriminative power. However, the validation results highlight that the 'stroma' class is difficult to classify. By adding more image patches, it may be possible to achieve better feature selection.

5.6 Conclusion

We have implemented and validated a computational method to identify and assess tissue image quality. This work shows that image descriptors of textures, intensities, entropy and etc, are strongly related to image quality. The analysis of these regions of interest can help to build a description of quality for the whole image.

Chapter 6

Automatic assessment of tissue micro-array quality

6.1 Introduction

As explained in Chapter 5, tissue samples are heterogeneous and can contain stromal tissue, clumps of cells, background staining and degradation of DNA. Pathologists look at the quality of the sample under a microscope to determine whether the sample should be selected for quantitative study. In order to assure the objectivity of tissue sample selection, it is important to develop automatic image quality assessment algorithms. The ideal tool should provide the pathologist with a recommendation to use a whole image or possibly only subareas. In Chapter 5, we have successfully built a texture based classification model to classify tissue regions of different quality. However, there are further challenges to overcome in order to automatically generate a quality report for images when the location, shape and size of different types of tissue are unknown. In order to solve the difficulties in automatical recognition of tissue types, this chapter tackles the problem from a different angle and builds a solution which recognises all the objects in an image. In addition, a set of TMA quality metrics are proposed and built based upon the recognition results.

The contributions of this chapter include: 1) presenting the first methodology

capable of automatically quantifying tissue quality; 2) implementing state-of-art classifiers to recognize different type of cells and tissue in microscopic images; 3) proposing an innovative quantitative microscopy method based on image features; and 4) validating the method against ground truth.

6.2 Literature Review

TMA quality control through quantification of tissue or cell type is innovative and there is little related work in the literature. The most closely related work is on quantification of certain types of tissue in a whole-slide image in histological images [70, 111, 109]. In Homeyer's work [70], an image texture based classification model was built to quantify regions contains necrotic tissue, viable tissue and background. That work partitioned the whole slide image (WSI) into image tiles of size 512×512 pixels ($120\mu m \times 120\mu m$), while a whole slide image often contained billions of pixels. Each image tile was labeled as either viable, necrotic or background. When an image tile contained multiple types of tissue, the tile was assigned to the one with largest area. That work combined the features from different scales and built a classifier for each tile. That classifier was then combined with a context classifier. That method analyzed each tile and the ratio between necrosis and viable tiles was computed as the amount of necrosis in the image. The division of image tiles led to errors in quantifying necrosis, but because whole slide images are very large[1], the error of this size was negligible.

In Mercan's work [109], sliding windows were used together with a bag-of-words model to locate diagnostically relevant regions of interest. In the reported experiments, a word was a 120×120 pixel patch cut from a whole slide image and a bag a 3600×3600 pixel window that contained 900 visual words. Image features used

[1]The biopsy is typically scanned at 40X magnification, and on average contain $90,000 \times 70,000$ pixels

for histology images were computed to represent each 'word' patch and a K-means algorithm used to cluster the 'words' as 50 or 100 clusters. Linear classifiers such as logistic regression and SVM were then trained to distinguish patches that belonged to regions of interest from the rest of the image.

In Doyle's work [46], a multi-resolution Bayesian classifier was proposed to detect regions of prostate cancer on digital biopsy slides. The whole slide image was decomposed into an image pyramid, and the regions that were considered as prostate cancer were examined at a higher spatial resolution. At each resolution level, an Adaboost classifier was built and three resolution scales used. Image features in each resolution were computed, such as Gabor filters, first-order-statistics, co-occurence, etc. At each image resolution, different types of image features become more relevant for discriminating between prostate cancer (CaP) and benign areas. In that work, the author also commented that the higher the resolution, the more difficult it was to build a classifier to distinguish CaP from a benign area.

Unfortunately, it is hard to use previous methods in our own study. This is because our images are much more smaller in size but higher in spatial resolution. The images in our study are taken at 60X magnification with 2048×2048 pixels and $120um \times 120um$ in size, equivalent to one tile of a whole slide image in Homeyer's work. At this spatial resolution, individual cells and their sub cellular components can be observed clearly, but there are only a few number of cells in one image. For a pathologist, the tissue types are recognized through evidence from cells and groups of cells. In this case, the errors caused by two or more tissues mixed together in a tile are significant. It is hard to assign a single label to an image tile when multiple types of cell exists in the same tile. It is also hard to use a 'bag-of-word' model because traditionally this model ignores location information of the object, which makes it hard

to calculate the proportion of each object in the image. In order to obtain location information, sliding windows have to be used. However, sliding windows are often computationally expensive and are hard to used in our case when multiple types of cell mix together [109].

In order to overcome the previous challenges, we have adopted an object proposal approach [157, 129, 87, 71]. When using an object proposal method for recognition, the image is first partitioned or segmented into different regions and they form a set of candidate objects. Instead of searching for an object from an image at every location and scale as with a sliding window method, this method partitions the image based on objects features, such as edge [157], geodesic segmentation [87], superpixels etc. This set of object proposals is then passed to a classification models.

6.3 Methodology

In this work, we considered five types of objects and they are: clumpy nuclei, degraded nuclei/nucleus, normal cancer nuclei/nucleus, stromal tissue and others. The type of 'others' defines the image patches that are hard to be characterised by eye. The definition for these object is based on information from DAPI channel. The aim is to recognise all these candidate objects in images.

An original analysis framework has been designed as follows, shown in Fig. 6.1. First of all, an object proposal is obtained based on segmentation of nuclei or nucleus in DAPI channel. At the location of a proposed object, image patches of a variety of sizes are used to obtain a texture and intensity feature description. Classification models are built for image patches and later for proposed objects as outlined below.

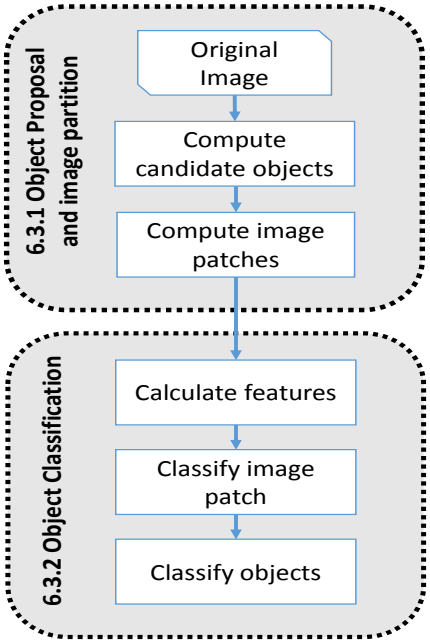


Figure 6.1: The overview of the framework.

6.3.1 Object Proposal and Image Partition

An object proposal is obtained based upon level set segmentation, and a partition of regions of the image is calculated. A level set model is used to generate an object proposal. The object proposal makes the assumption that each cell has a visual nuclei that is shown in a DAPI channel of the image. Level set segmentation is explained in Chapter 2. The images are firstly partitioned into large regions based on the shape of proposed object. A watershed tessellation is used for the purpose. As shown in Fig. 6.2 (b), evidence of cell nuclei naturally partitions an image into different regions (Fig. 6.2 (c)). After these regions have been obtained, image patches are located to cover these regions. By this way, one candidate object is represented by a number of image patches of a variety of sizes. The image patches are also characterised into five different group, which are: 'clumpy', 'degradation', 'normal', 'stroma' and 'others'. The aim is to recognise complete objects in each partition based on information from the image patches.

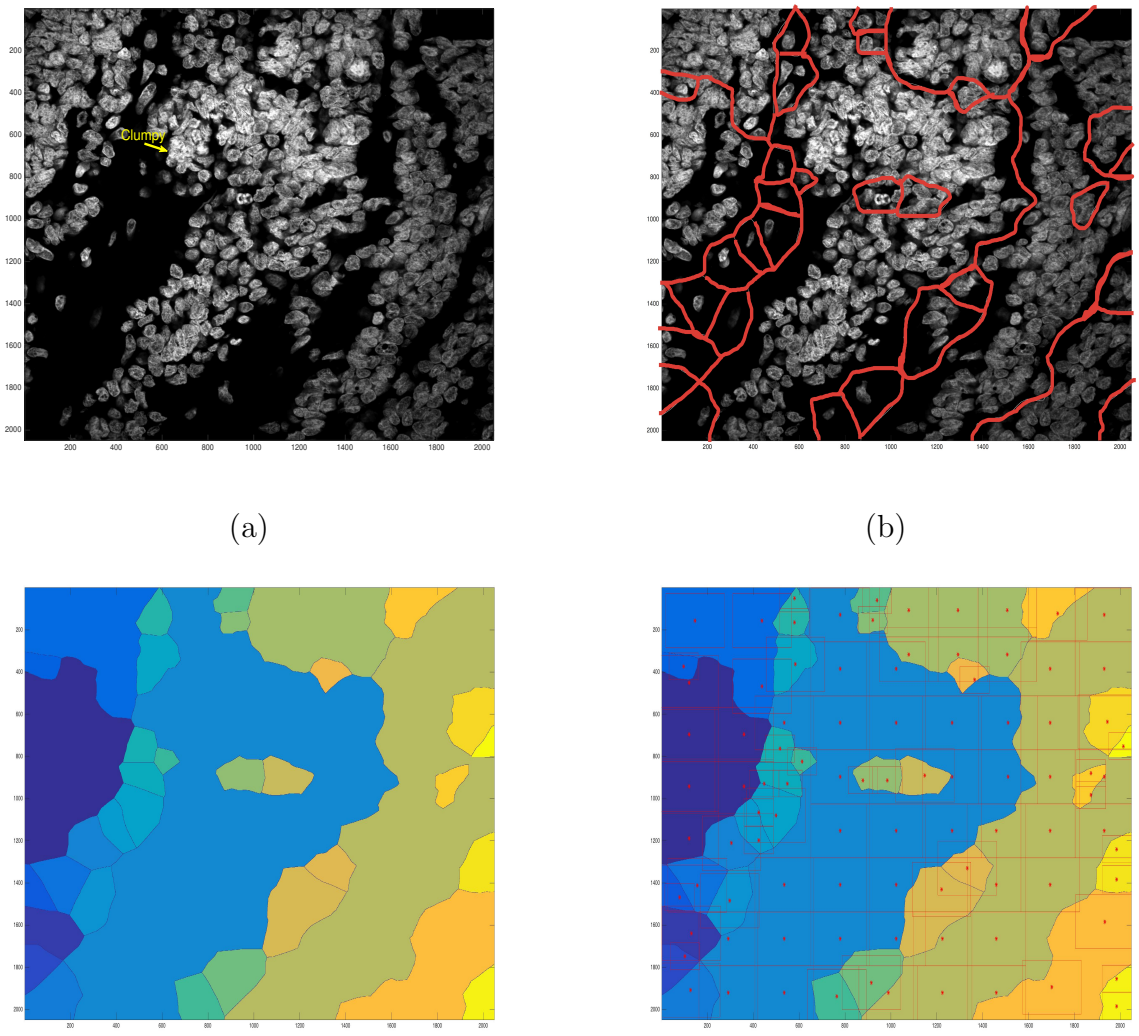


Figure 6.2: An example of the object proposal of an TMA image. (a)The original DAPI channel of the TMA image. This image shows a large amount of cell nuclei form clusters. (b) The partition of the image based on DAPI channel. (c) The image regions obtained from image partition. (d) An example of a variety of sized image patches.

Under this framework, the choice of image patches is important for analysis. The objects that we want to recognise are different in size and shape. For example, the 'clumpy' tissue regions where cells are all clumped together, can be any size that is bigger than two cells. Sometimes, the cell nuclei seems to be degraded and the nuclei shows no boundaries between each other. In this case, the object is both clumpy and degraded, but the image patches of the object shows sign of degradation instead of clumpy. So the object is considered as degradation instead of clumpy.

In order to be robust to objects with different sizes and shapes, we adopted image patches with predefined size (see Fig. 6.2 (d)). The method of calculating the size of image patches is described in A. If the object is smaller than the largest image patch, the size that is most close to the object size is adopted as the size of the image patch. The image patches in this study are centered at the geometric centroid of the candidate object. If the object is larger than the largest patch size, several patches with the largest size are used. The location of these patches are calculated to cover the object. In another word, each region of an object is filled with these square tiles. If the region is larger than a certain size, several tiles are computed to fit the region. The location and the size of the tiles are calculated for each region. These regions form the image features for this region. The task of classification is to successfully classify these regions.

6.3.2 Object Classification

There are two steps in the classification model. The first step classifies each patch into different classes based on texture information and the second step considers contextual information from other patches of the same object. The prediction of the

object is produced by considered both patch detection and it's contextual information. In order to generate classification results for each object, the follow procedure is followed.

Image Patch Classification A multi-class random forest is used to learn the image patches associated with each object type. In order to build the classifier for each patch, image features are computed, as explained in Chapter 5. The features are selected using sequential feature selection (Details in Chapter 5 Section 5.4.1). The features of each patch are computed and selected with forward sequential feature selection. The multi-class random forest is then built to classify the image patches.

Image Patch Level Prior The classification of an object is a combination of a prior probability indicating the likelihood of an object observed under a certain size of image patch and the likelihood of an image patch belong to a certain class. Under this framework, an object that is large often has multiple patches and is observed with a large image patch. Smaller object are likely to be observed with a smaller image patch. There is a natural prior distribution of each classes at different image patch levels. This prior was determined empirically in the experiments.

Contextual Information Once classification of image patches is obtained, for each object, an average voting procedure from all the image patches was computed. This voting procedure is an area weighted voting scheme. The final classification score is based upon the average from all the patches of this object. The weighting factors mainly control the voting power of the patches that belong to the same object. The

weighting factor is calculated based upon the ratio between the area of the partition space which is within the image patch and the area of the image patch.

6.4 Experiments

6.4.1 Object classification

The images in this experiment were selected and labelled by a pathologist. The pathologist defined image patches belong to a certain type of object. In the training step, 24 images were randomly selected by the pathologist and 6 of each contained clumpy, degraded, normal and stromal tissue. This design allowed for the image patches to be of variety of sizes. This experiment choose to use three different scales. The image patch sizes were 64×64 pixels (named Level 2), 128×128 pixels (named Level 1) and 256×256 pixels (named Level 0). These sizes were determined empirically. From the images, 658 level 2,985 level 1 and 1224 level 0 image patches were taken using the method explained previously, and manually labelled by a pathologist.

The distribution of labels are shown in Fig. 6.3. The distribution of each class are highly skewed. For example, the number of observations of clumpy patches are more than that of any other type in Level 1 image patches. The larger window tends to contain more clumpy nuclei while for a smaller window there tends to be patches that are harder to characterise by eye (these patches are labelled as 'other'). The unbalanced distribution of each type of patch indicates that there is a natural prior for the image patches. Under different window sizes, different objects are more likely to be observed.

A resampling method was used to produce balanced data, build a classifier and then apply the prior. This method produced a balanced dataset for classification. The pathologist manually labelled more data for the group that had a few datapoints. Due

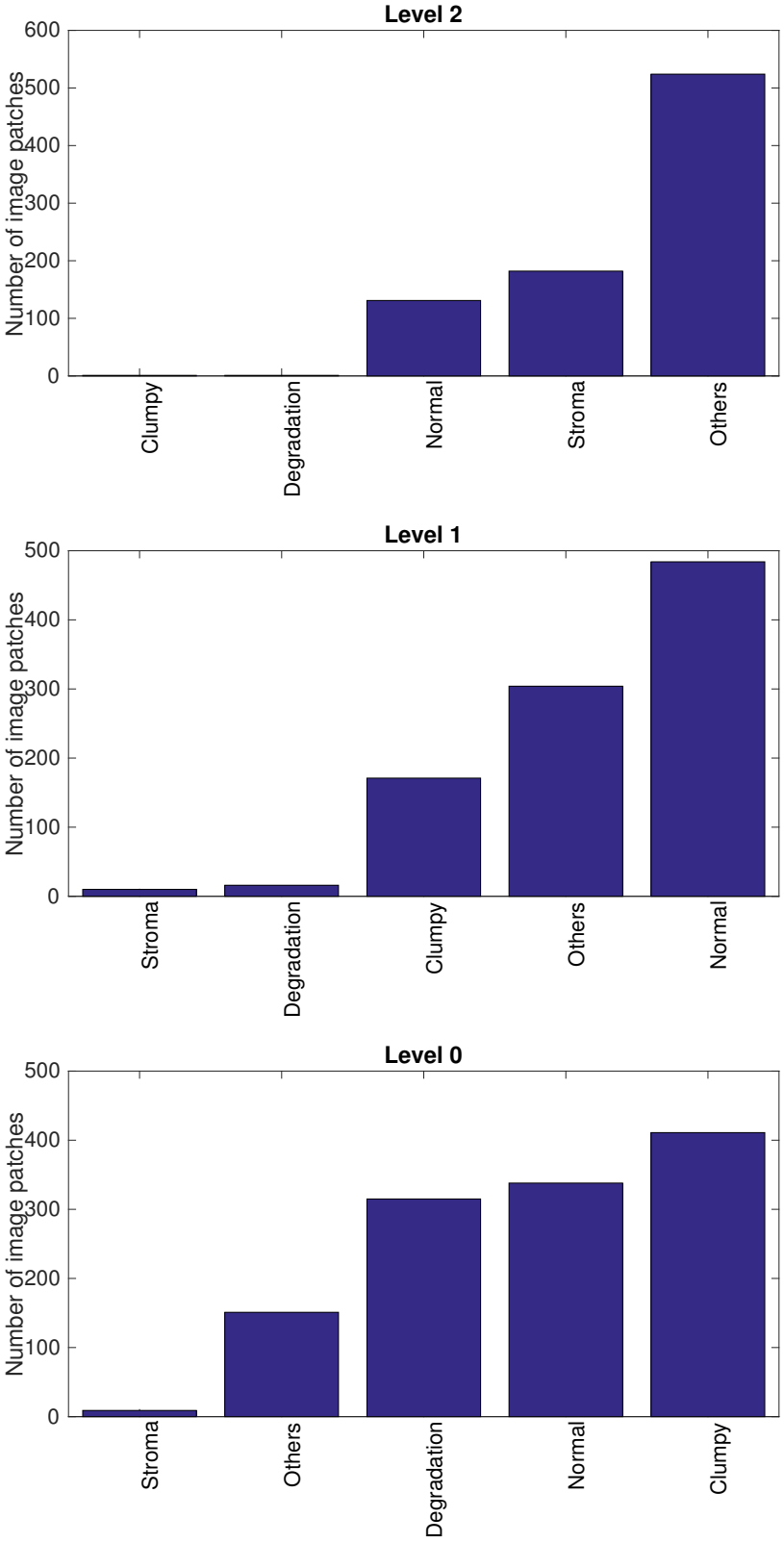


Figure 6.3: Examples of image patches of tissue samples.

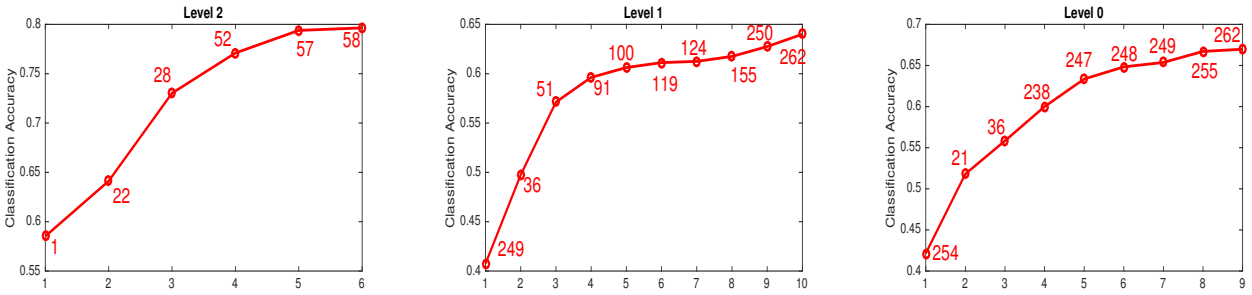
Features	Vector (Level 0 & 1)	Length (Level 0 & 1)	Vector (Level 2)	Length (Level 2)
SCN	1-242	242	1-52	52
Auto-correlation	242-244	3	52-54	3
Intensity	245-248	4	55-58	4
Major Edge Density	249	1	59	1
Entropy	250-253	4	60-63	4
Correlation	254	1	64	1
MSER	255-260	6	65-68	6
HoG	261-262	2	69-70	2

Table 6.1: The break down of the complete descriptor.

to the limitation of the number of images available, some of the datapoints were also randomly up-sampled. The group with a large number of datapoints, was resampled to reduce the number of included samples.

The complete image feature vector for each level is shown in Table 6.1. The features that were selected using the sequential feature selection are shown in Fig. 6.4. For level 2 image patches, SCN features and the skewness and kurtosis of intensity distribution were found to be discriminative. For level 1 image patches, SCN features, MED, Entropy and HoG are important features. For level 0, correlation, intensity skewness and intensity kurtosis, MSER shape and HoG were found to be discriminative.

A multi-class decision tree classification model was learnt from the image patches (Details in Chapter 5). The trees were grown to a sufficient depth to capture small samples. The features were selected using sequential feature selection (Details in Chapter 5 Section 5.4.1). The features of each patch were computed and sequential feature selection was used. After the model was built, the model was cross validated using leave-one-out cross-validation. In this study, 20% of the data were left out and the results are shown in Fig. 6.2.



(a) Level 2 (b) Level 1 (c) Level 0

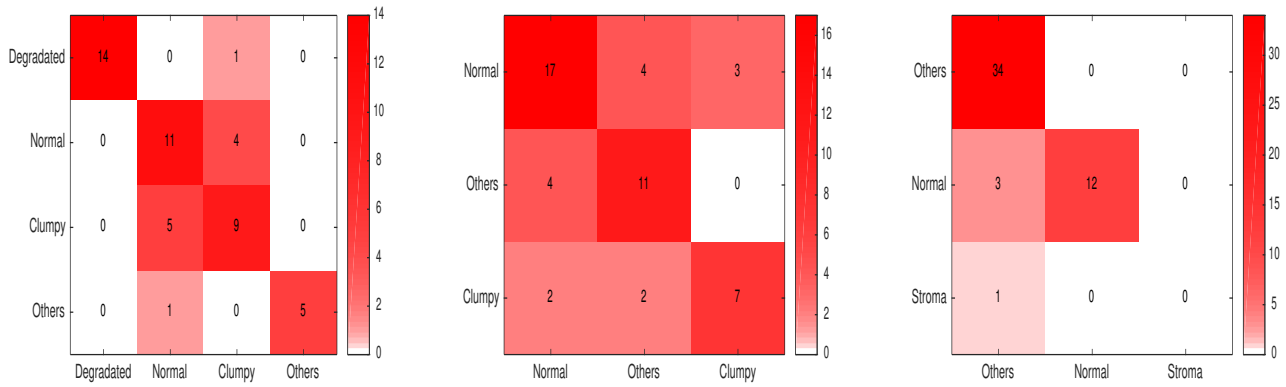
Figure 6.4: Sequential selection of discriminative image features. The accuracy of RFs vary with different their levels and y-axis are different. The numbers are corresponding to the location in vector in Table 6.1.

Model	No.Trees	leave-one-out (%)
RF Level 0	200	65.4±0.4
RF Level 1	200	60.3±0.5
RF Level 2	200	78.9±0.6

Table 6.2: Cross validation of random forest model.

The prior for image patches was tuned to reflect the proportion of each class in Fig. 6.3. In this experiment, the class with very small number of patches were given a smaller value. For example, for level 2, the likelihood that a patch belong to 'clumpy' or 'degradation' class is set to 0, because under this framework, it's almost impossible that 'clumpy' object is observed using level 2 patches. The likelihoods of an level 2 patch belong to 'normal', 'stroma' or 'others' are more than 0 and the sum of them is equal to 1. The prior was then applied to the classification of candidate objects. After applying the prior, the classification of object achieved good performance. The classification of image patches are improved, as shown in Fig. 6.5.

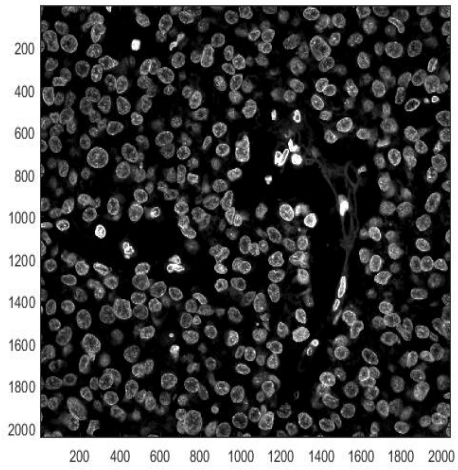
Three independent images are used to test the classification of image patches after applying the prior adjustment. The images were partitioned into image patches. 50 level 2, level 1 and level 0 image patches were randomly selected from these patches

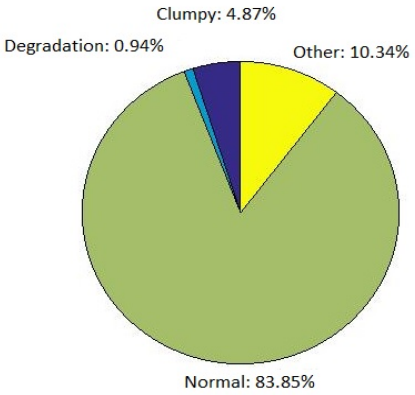


(a) Level 0 (b) Level 1 (c) Level 2

Figure 6.5: Evaluation of the classification on independent dataset. The result is showing the performance of the classifier when it is used to classify the data. (a) Classification results of Level 0. (b) Classification results of Level 1. (c) Classification results of Level 2.

and classified. The results of the classification is compared against ground truth. The confusion matrices are shown in Fig. 6.5. Due to the highly skewness of the image patches in each level, the test dataset are also skewed. In this experiment, it was difficult to find image patches with stromal cells in it. In this dataset, there was few stromal cells observed. In Fig. 6.5 (a), 14 'degraded' image patches were successfully classified as 'degraded', while 1 'degraded' image patch was classified as 'clumpy'. 11 'normal' image patches were successfully classified as 'normal', while 4 'normal' image patches were classified as 'clumpy'. 9 'clumpy' image patches were classified as 'clumpy', while 5 were classified as 'normal'. 5 'others' were classified as 'others' and 1 was classified as 'normal'. This results shows the performance of level 0 classifier is very good. For level 1 and level 2 classifier, their performance are also very good on the testing dataset, as shown in Fig. 6.5 (b) and (c).





(a) (b)

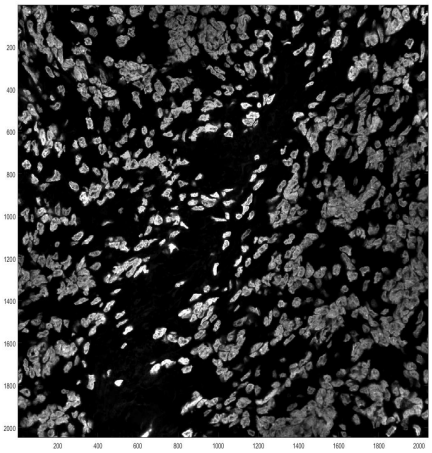
Figure 6.6: An examples of quantification tissue quality. (a) The original image for analysis. (b) The proportion of each type of tissue in this image found by the algorithm.

6.4.2 Quantification of Tissue with Different Quality

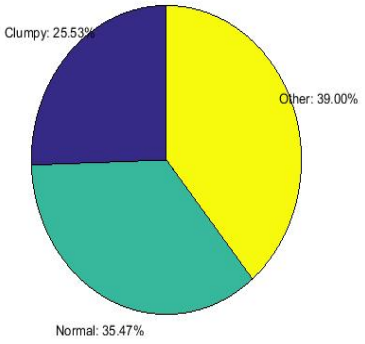
The proportion of each type of tissue in the image were computed. The area of each type of cell and tissue can be obtained from the classification model and for each image, a pie chart was used to show the proportion of objects found in the image. As shown in the example Fig. 6.6, the proportion of each type of cells are presented in a pie chart. More examples are shown in Fig. 6.7.

6.4.3 Image Quality Control

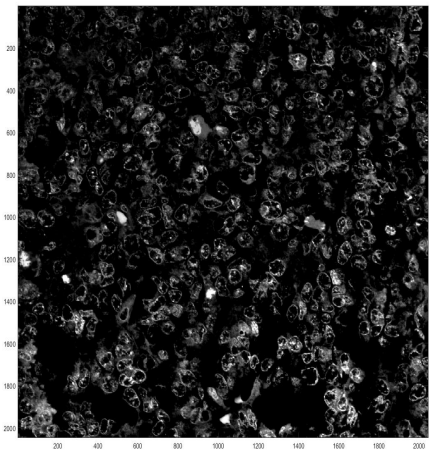
The proportion of each of tissue in the image can form a quality metric for the image. In order to control the quality of the images, it is important to build a recommendation algorithm. In order to build tissue quality control algorithm, a random forest mode was used to classify the images that should be excluded from the analysis. The proportion of each type of tissue was calculated. The proportion of each type of tissue was used as features to build a classification model. The leave-one-out cross-



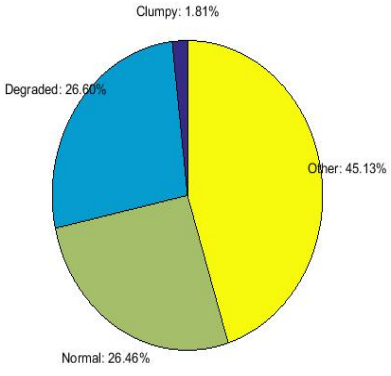
(a)



(b)



(c)



(d)

Figure 6.7: More examples of tissue images. (a) an example image. (b) the proportion of each type of tissue in image (a) found by the algorithm. (c) further example image. (d) The proportion of each tissue type found by the algorithm from image (c).

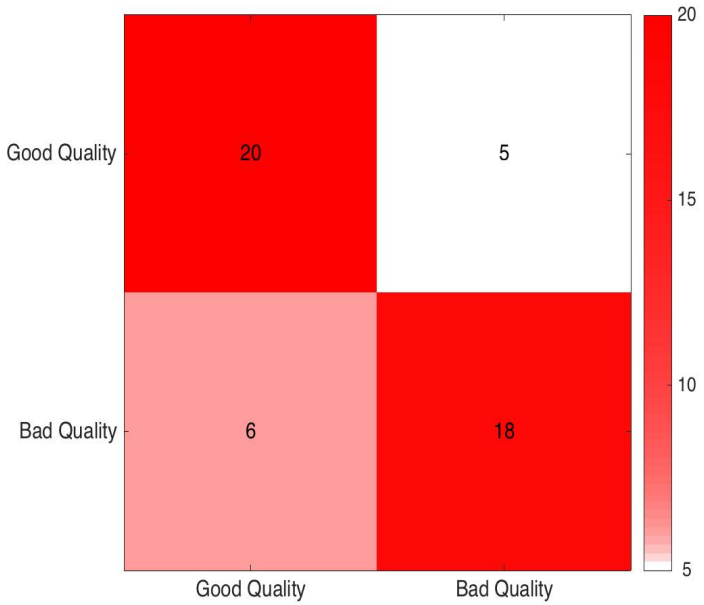


Figure 6.8: Test results reported as a confusion matrix. The figure shows that 20 good quality images are classified as good quality while 18 bad quality are classified as bad quality. In addition, 6 images that are good quality, but classified as bad quality. 5 images are bad quality and classified as good quality.

validation results shows 79% percent of accuracy. In addition, an independent test of images shows that the sensitivity is 77% and specificity is 78%[2], as shown in Fig. 6.8.

6.4.4 Correlation with Nuclei Segmentation

The segmentation of nuclei as explained in Chapter 4 has a better accuracy when the tissue image is of good quality and vice versa. In previous research [23], the selection of informative cells was based on the area of nuclei, which was between 1500 pixels and 8000 pixels [3]. In our research, we defined the informative nuclei ratio (INR),

[2]The sensitivity is calculated as: $Sensitivity = \frac{TruePositive}{TruePositive+FalseNegative}$ and the speicificity is calculated as $Specificity = \frac{TrueNegative}{TrueNegative+FalsePositive}$.

[3]When the tissue image is difficult to segment, the segmentation algorithm generate errors and results in under-segmentation or over-segmentation. In addition, some background noise and out-of-focus nuclei might be detected. In order to solve these problem, only the cells that has nuclei in a certain size are used for quantitative analysis.

which is the ratio between the total area of the informative nuclei and the total area of segmented tissue to quantify the measurability of an image.

In this experiment, 60 images were randomly selected. The experimental INR is computed for these images and the image quality metric of them are computed. The experimentally calculated INR was based on the automatic image segmentation algorithm in Chapter 4. A random forest regression model was built to fit the quality metric with the experimental INR. The image tissue quality metric that we have defined, have been used to build random forest regression model, which predict the INR for an image.

In an independent test set of 40 images, the predicted INR was compared with the experimentally found INR. The predicted value shows great correlation with the experimental INR. As shown in Fig. 6.9, Bland-Altman test is used to study the difference between the predict INR and the experimental INR. It shows that the experimental INR the predict INR are close to each other.

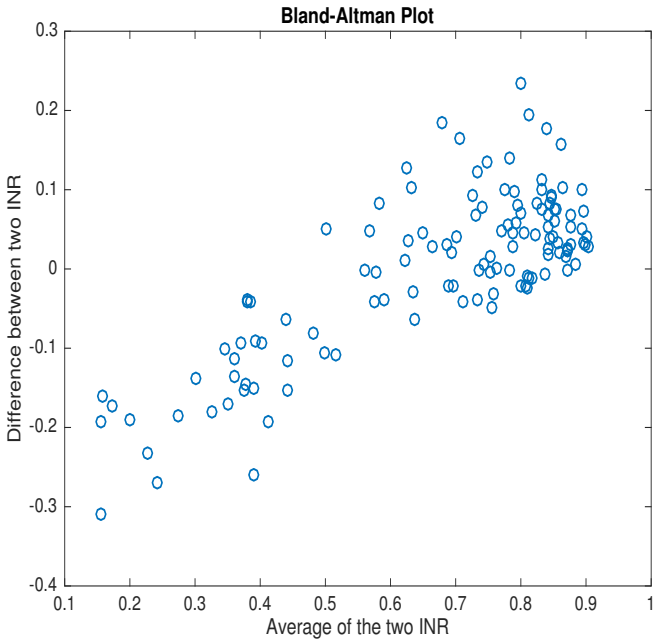


Figure 6.9: The Bland-Altman test of INR by predicted computer and the INR calculated in experiment.

Tissue Type	Mean Error (%)	Std (%)	Mean Square Error (%)
Clumpy	+3.2	6	5.3
Normal	+2.3	6	4.2
Others	-7	5	7.1
Degradation	+3.4	6.2	4.7
Stroma	-1.9	0.9	1.8

Table 6.3: The average error of the algorithm against ground truth.

6.5 Validation of Tissue Quantification

Three new images were randomly selected by the pathologist and manually labelled. The results are compared with the algorithm. As shown in Table. 6.5, the agreement between the manual label and algorithm label is good (see also Fig. 6.10). The '+' sign in the table is indicating the algorithm over estimates the proportion of the tissue type and the '−' sign is indicating the algorithm is under estimate the proportion.

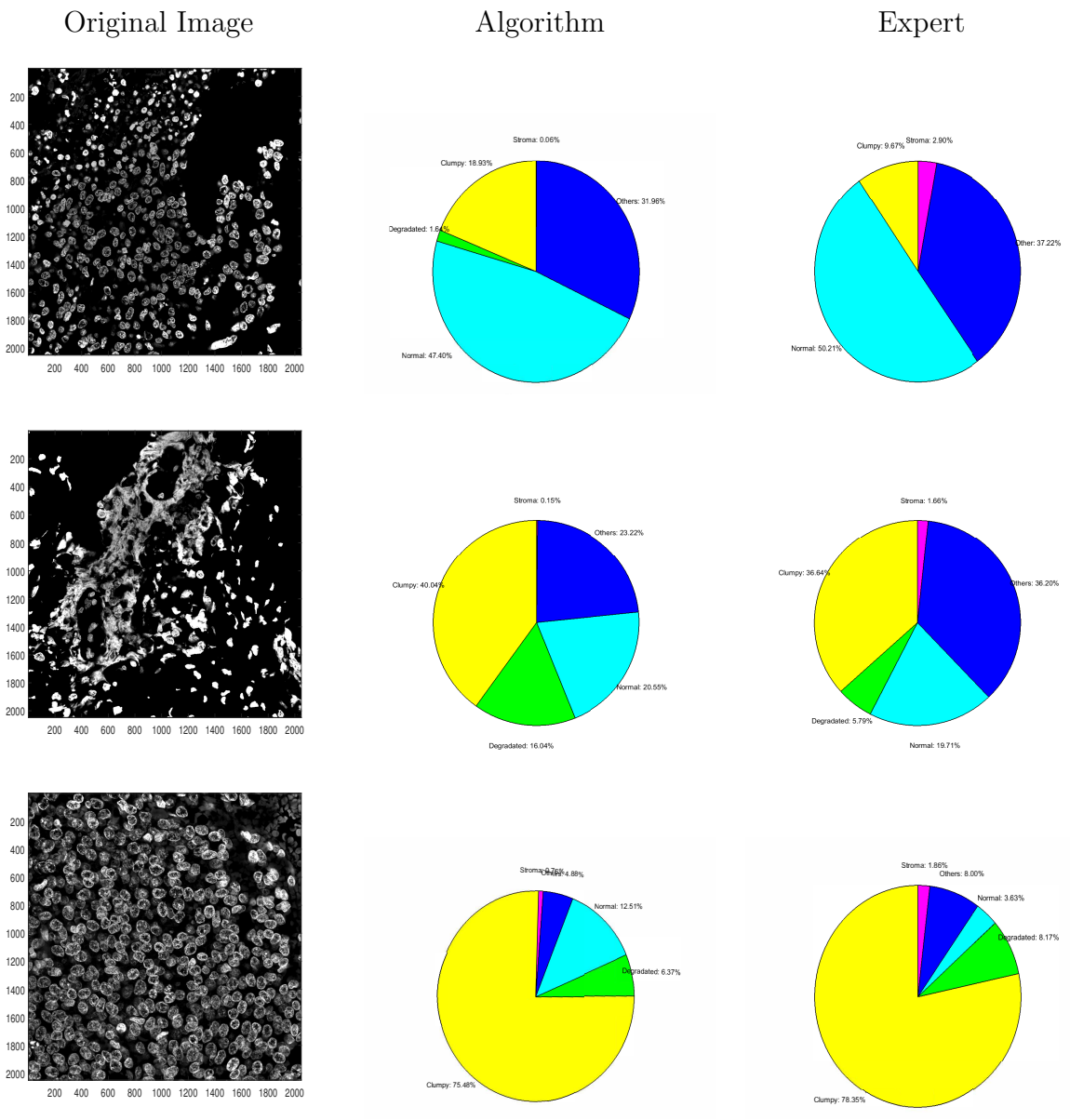


Figure 6.10: The proportion of each object that are labelled by computer is compared with that of a pathologist. The pie chart of the left hand side is showing the proportion of each type of object identified by computer algorithm. The right hand side is showing the proportion of each type of object identified by a pathologist.

6.6 Discussion

Tissue Quality The metrics of tissue type in this analysis mainly focused on clumpy, degradation and stromal tissue. There might be other types of tissue and artefacts that may also affect the quantification results, such as dust etc. In the image data set available for our work, there were not enough images to study dust. However, if more images become available, it would be possible to train a more complicated classification model and recognition these different tissue types.

Object classification The main challenge of object recognition in this study was to build a practical classifier. Due to the relatively small number of images available with bad quality, the estimation of the image patch prior and selection of discriminative features was suboptimal. Including more images would improve the accuracy of recognition. The number of images with stromal cells in it are really small, and it is hard to study stromal tissue due to the limitation of images. However, it doesn't change the message of this chapter that such a framework is capable of recognise tissue with different quality. Further improvement can be made when more images become available.

6.7 Summary

This chapter has proposed an innovative framework to quantify tissue image quality through recognition of the tissue types. Under this framework a tissue quality metrics was built and validated in a pilot study.

Chapter 7

Conclusions and Future Work

This work explores the quantification of immuno-fluorescence tissue microscopic images and implements an analysis pipeline, which includes the four main contributions of this thesis.

The first contribution builds and validates a fluorescence microscopic images using retrospective method. This work is really important for taking an accurate measurement from the image.

The second contribution develops an original automatic cell segmentation for the fluorescent tissue images. This work facilitates the computation of quantitative features for each cell, and was published in a high-impact journal paper.

The third contribution investigates the textures of different tissue quality. When there are different types of tissue (related to their quality) in an image, image features are computed to build an descriptor. Discriminative features were found and classification models were implemented to classify the images. This work was presented in a high-impact international workshop.

The fourth contribution constructs a framework to analyse tissue quality. Tissue

quality control is one of the most important part of TMA analysis. An original tissue quality control algorithm is built to assist a pathologist in selecting samples in an unbiased manner. This work is currently in preparation for a high impact journal article.

In conclusion, this thsis builds an initial pipeline to analyse immuno-flourescence tissue microarray images. The quantitative analysis of tissue microarray images can assist pathologists in analysing their images and supports the study of cancer biomarkers.

7.1 Future Work

In this quantitative pipeline, the encountered limitations have already been discussed in each of the chapter. The improvements of each component in the pipeline are also highlighted in each chapter. However, there are further challenges arising from this work. There are new research directions and interests based on this thesis, such as translational work and deep learning. They are explained in details.

7.1.1 Translational Work

One direction for future research is to take this algorithm forward to clinical studies. A natural extension is to investigate correlation between patterns in Ewing Sarcoma images and the clinical outcomes. Statistical models [23, 6] can be built to analyse the heterogeneity of tumour patterns and to study the patient outcome data.

7.1.2 Deep Learning

Another direction to take this framework forward is to study deep convolutional neural networks (CNN). In Chapter 6, a framework is proposed to study tissue quality

through the recognition of tissue types. In future, CNNs may improve recognition accuracy, as it has already shown great potentials in computer vision applications [88, 91, 140, 99]. It will be interesting to use CNNs to mine the subtle changes in cells which are related to their underlying biology. Using current knowledge, it is not yet clear, and how to adopt CNNs within the tissue image analysis framework.

In order to adapt the CNNs method, a large volume of imaging data is normally necessary [88, 91]. The volume of images obtained from patients is normally limited and more images needs to be acquired. It is still unclear how to improve data efficiency of CNN so that it is possible to apply this promising technique on a much smaller dataset.

Appendix A

```
% calculate the boudingbox correlated the center
% and patchsize
% input wtsregion : is a region in the paritions of an image
% bdSz: is the maximum size of the image pathces.

function nbox = calculate_boundingbox(wtsregion,bdSz)
%% fit detection boxes

stats= regionprops(wtsregion,'BoundingBox');
boundingXY  = stats.BoundingBox;
if nargin

    bdSz = [8,8];
end
thisbox(1).center =[];
thisbox(1).patchSz = [];
% fit the patchSz that are smaller than patchSz;
   x0 = boundingXY(1);
    x =  boundingXY(3);
    y0 = boundingXY(2);
    y = boundingXY(4);
 %
    n = round(log2(x));
    m = round(log2(y));
%
    if n>=bdSz(1)&&m>=bdSz(2)
        num_x = round(x/2^bdSz(1));
        num_y = round(y/2^bdSz(2));
        p = 1;
        for i = 1:num_x
            for j = 1:num_y
            thisbox(p).center = [round(x0+x/num_x/2+x/num_x*(i-1)),round↙
(y0+y/num_y/2+y/num_y*(j-1))];
            thisbox(p).patchSz = [2^bdSz(1),2^bdSz(2)];
            p = p+1;
            end
        end
    elseif or(n<6,m<6)
        thisbox.center = [round(x0+x/2),round(y0+y/2)];
        thisbox.patchSz = [64,64];
    elseif or(n<bdSz(1),m<bdSz(2))&&n>=6&&m>=6
        [M I] = min([n,m]);
```

```
        num_x = round(x/2^M);
        num_y = round(y/2^M);
        p = 1;
        for i = 1:num_x
            for j = 1:num_y
                thisbox(p).center = [round(x0+x/num_x*i-2^(M-1)),round(y0+y/num_y*j-2^(M-↙
1))];
                thisbox(p).patchSz = [2^M,2^M];
                p = p+1;
            end
        end
    else
        fprintf('What is the size of the bouding thisbox????');
    end

%%
sz = size(thisbox);
if sz(1)>1
   for i = 1:length(thisbox)
           temp = thisbox(i).center;
           center_matrix(temp(2),temp(1)) = wtsregion(temp(2),temp(1))*i;

      end
else
      temp = thisbox(1).center;
      center_matrix(temp(2),temp(1)) = 1;
end

        [v] = unique(center_matrix);
        v1 = v(v~=0);
       nbox = thisbox(v1);
end
```

Bibliography

[1] et al. Adiga, Umesh. High-throughput analysis of multispectral images of breast cancer tissue. *Image Processing, IEEE Transactions on*, 15(8):2259–2268, August 2006.

[2] Manoj Aggarwal, Hong Hua, and Narendra Ahuja. On cosine-fourth and vignetting effects in real lenses. In *Computer Vision, 2001. ICCV 2001. Proceedings. Eighth IEEE International Conference on*, volume 1, pages 472–479. IEEE, 2001.

[3] Amir Ahmad and Lipika Dey. A k-mean clustering algorithm for mixed numeric and categorical data. *Data & Knowledge Engineering*, 63(2):503–527, 2007.

[4] Yousef Al-Kofahi, Wiem Lassoued, William Lee, and Badrinath Roysam. Improved automatic detection and segmentation of cell nuclei in histopathology images. *Biomedical Engineering, IEEE Transactions on*, 57(4):841–852, 2010.

[5] Ash A Alizadeh, Michael B Eisen, R Eric Davis, Chi Ma, Izidore S Lossos, Andreas Rosenwald, Jennifer C Boldrick, Hajeer Sabet, Truc Tran, Xin Yu, et al. Distinct types of diffuse large b-cell lymphoma identified by gene expression profiling. *Nature*, 403(6769):503–511, 2000.

[6] Marco Archetti, Daniela A Ferraro, and Gerhard Christofori. Heterogeneity for igf-ii production maintained by public goods dynamics in neuroendocrine pan-

creatic cancer. *Proceedings of the National Academy of Sciences*, 112(6):1833–1838, 2015.

[7] Salim Arslan, Tulin Ersahin, Rengül Cetin-Atalay, and Cigdem Gunduz-Demir. Attributed relational graphs for cell nucleus segmentation in fluorescence microscopy images. *Medical Imaging, IEEE Transactions on*, 32(6):1121–1131, 2013.

[8] Lidia Auret and Chris Aldrich. Empirical comparison of tree ensemble variable importance measures. *Chemometrics and Intelligent Laboratory Systems*, 105(2):157–170, 2011.

[9] Georgios Babaloukas, Nicholas Tentolouris, Stavros Liatis, Alexandra Sklavounou, and Despoina Perrea. Evaluation of three methods for retrospective correction of vignetting on medical microscopy images utilizing two open source software tools. *Journal of microscopy*, 244(3):320–324, 2011.

[10] Chris Bakal, John Aach, George Church, and Norbert Perrimon. Quantitative morphological signatures define local signaling networks regulating cell morphology. *science*, 316(5832):1753–1756, 2007.

[11] Christopher F. Batten. Autofocusing and Astigmatism Correction in the Scanning ElectronMicroscope. *[http://www.csl.cornell.edu/ cbatten/pdfs/batten-image-processing-sem-ucthesis2000.pdf]*, pages 1–89, August 2000.

[12] BBC. Half of UK population 'will get cancer in lifetime'. Technical report, [http://www.bbc.co.uk/news/health-22796220], 2013.

[13] Yoshua Bengio, Jean-François Paiement, Pascal Vincent, Olivier Delalleau, Nicolas Le Roux, and Marie Ouimet. Out-of-sample extensions for lle, isomap, mds, eigenmaps, and spectral clustering. *Advances in neural information processing systems*, 16:177–184, 2004.

[14] John D Benson, Ying-Nan P Chen, Susan A Cornell-Kennon, Marion Dorsch, Sunkyu Kim, Magdalena Leszczyniecka, William R Sellers, and Christoph Lengauer. Validating cancer drug targets. *Nature*, 441(7092):451–456, 2006.

[15] Jan-Philip Bergeest and Karl Rohr. Efficient globally optimal segmentation of cells in fluorescence microscopy images using level sets and convex energy functionals. *Med Image Anal*, 16(7):1436–1444, October 2012.

[16] Christopher M Bishop. Pattern recognition. *Machine Learning*, 2006.

[17] Avrim L Blum and Pat Langley. Selection of relevant features and examples in machine learning. *Artificial intelligence*, 97(1):245–271, 1997.

[18] Yuri Boykov and Gareth Funka-Lea. Graph Cuts and Efficient N-D Image Segmentation. *International Journal of Computer Vision*, 70(2):109–131, November 2006.

[19] Yuri Y. Boykov and Marie-Pierre Jolly. Interactive graph cuts for optimal boundary and region segmentation of objects in n-d images. *Computer Vision, 2001, ICCV, 2001. Proceedings. Eighth IEEE*, 1:105–112, 2001.

[20] Leo Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.

[21] Donal J Brennan, Darran P O'Connor, Elton Rexhepaj, Fredrik Ponten, and William M Gallagher. Antibody-based proteomics: fast-tracking molecular diagnostics in oncology. *Nat Rev Cancer*, 10(9):605–617, September 2010.

[22] Joan Bruna and Stéphane Mallat. Invariant scattering convolution networks. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 35(8):1872–1886, 2013.

[23] Claudia Bühnemann, Simon Li, Haiyue Yu, Harriet Branford White, Karl L Schäfer, Antonio Llombart-Bosch, Isidro Machado, Piero Picci, Pancras CW

Hogendoorn, Nicholas A Athanasou, et al. Quantification of the heterogeneity of prognostic cellular biomarkers in ewing sarcoma using automated image and random survival forest analysis. *PloS one*, 2014.

[24] Robert L Camp, Gina G Chung, and David L Rimm. Automated subcellular localization and quantification of protein expression in tissue microarrays. *Nature medicine*, 8(11):1323–1328, 2002.

[25] Pablo G Cavalcanti and Jacob Scharcanski. Automated prescreening of pigmented skin lesions using standard cameras. *Computerized Medical Imaging and Graphics*, 35(6):481–491, 2011.

[26] Tony F Chan, Luminita Vese, et al. Active contours without edges. *Image processing, IEEE transactions on*, 10(2):266–277, 2001.

[27] Hang Chang, Quan Wen, and Bahram Parvin. Coupled segmentation of nuclear and membrane-bound macromolecules through voting and multiphase level set. *Pattern recognition*, 48(3):882–893, 2015.

[28] Cheng Chen, Houqiang Li, Xiaobo Zhou, and Stephen TC Wong. Graph Cut Based Active Contour for Automated Cellular Image Segmentation in High Throughput Rna Interface (rnai) Screening. pages 69–72, 2007.

[29] Xiang Chen, Meel Velliste, Shmuel Weinstein, Jonathan W Jarvik, and Robert F Murphy. Location proteomics: building subcellular location trees from high-resolution 3D fluorescence microscope images of randomly tagged proteins. In *Biomedical Optics 2003*, pages 298–306. International Society for Optics and Photonics, 2003.

[30] Xiaowei Chen, Xiaobo Zhou, and Stephen TC Wong. Automated segmentation, classification, and tracking of cancer cell nuclei in time-lapse microscopy. *Biomedical Engineering, IEEE Transactions on*, 53(4):762–766, 2006.

[31] Yung-Fu Chen, Po-Chi Huang, Ker-Cheng Lin, Hsuan-Hung Lin, Li-En Wang, Chung-Chuan Cheng, Tsung-Po Chen, Yung-Kuan Chan, and John Y Chiang. Semi-automatic segmentation and classification of pap smear cells. *Biomedical and Health Informatics, IEEE Journal of*, 18(1):94–108, 2014.

[32] Koei Chin, Carlos Ortiz de Solorzano, David Knowles, Arthur Jones, William Chou, Enrique Garcia Rodriguez, Wen-Lin Kuo, Britt-Marie Ljung, Karen Chew, Kenneth Myambo, et al. In situ analyses of genome instability in breast cancer. *Nature genetics*, 36(9):984–988, 2004.

[33] EH Cho and SJ Lockett. Calibration and standardization of the emission light path of confocal microscopes. *Journal of microscopy*, 223(1):15–25, 2006.

[34] William C S Cho. MicroRNAs: potential biomarkers for cancer diagnosis, prognosis and targets for therapy. *Int J Biochem Cell Biol*, 42(8):1273–1281, August 2010.

[35] Luís Pedro Coelho, Aabid Shariff, and Robert F Murphy. Nuclear segmentation in microscope cell images: a hand-segmented dataset and comparison of algorithms. In *Biomedical Imaging: From Nano to Macro, 2009. ISBI'09. IEEE International Symposium on*, pages 518–521. IEEE, 2009.

[36] Ge Cong and Bahram Parvin. Model-based segmentation of nuclei. *Pattern Recognition*, 33(8):1383–1393, 2000.

[37] Christian Conrad, Holger Erfle, Patrick Warnat, Nathalie Daigle, Thomas Lörch, Jan Ellenberg, Rainer Pepperkok, and Roland Eils. Automatic identification of subcellular phenotypes on human cell arrays. *Genome research*, 14(6):1130–1136, 2004.

[38] Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine learning*, 20(3):273–297, 1995.

[39] Navneet Dalal and Bill Triggs. Histograms of oriented gradients for human detection. In *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, volume 1, pages 886–893. IEEE, 2005.

[40] Stephen W.Paddock. Thomas J. Fellers. Michael W. Davidson. Basic concepts.

[41] Armand de Gramont, Sarah Watson, Lee M Ellis, Jordi Rodón, Josep Tabernero, Aimery de Gramont, and Stanley R Hamilton. Pragmatic issues in biomarker evaluation for targeted therapies in cancer. *Nature Reviews Clinical Oncology*, 12(4):197–212, 2015.

[42] Sébastien Derivaux, Sébastien Lefevre, Cédric Wemmert, and Jerzy Korczak. On machine learning in watershed segmentation. In *Machine Learning for Signal Processing, 2007 IEEE Workshop on*, pages 187–192. IEEE, 2007.

[43] Ramón Díaz-Uriarte and Sara Alvarez De Andres. Gene selection and classification of microarray data using random forest. *BMC bioinformatics*, 7(1):1, 2006.

[44] Richard Dosselmann and Xue Dong Yang. No-Reference Image Quality Assessment using Level-of-Detail. Technical report, University of Regina, 2011.

[45] Colin Doutre and Panos Nasiopoulos. Fast vignetting correction and color matching for panoramic image stitching. In *Image Processing (ICIP), 2009 16th IEEE International Conference on*, pages 709–712. IEEE, 2009.

[46] Scott Doyle, Michael Feldman, John Tomaszewski, and Anant Madabhushi. A boosted bayesian multiresolution classifier for prostate cancer detection from digitized needle biopsies. *Biomedical Engineering, IEEE Transactions on*, 59(5):1205–1218, 2012.

[47] Alex Eccleston and Ritu Dhand. Signalling in cancer. *nature insight*, 441(7092):423, May 2006.

[48] Seth Eckhouse, Grant Lewison, and Richard Sullivan. Trends in the global funding and activity of cancer research. *Molecular oncology*, 2(1):20–32, 2008.

[49] M Economou, L Schöni, and C Hammer. Proper paraffin slide storage is crucial for translational research projects involving immunohistochemistry stains. *... and translational ...*, 2014.

[50] Nature Education. Cell signalling. *Scitable. [http://www.nature.com/scitable/topicpage/cell-signaling-14047077]*, 2015.

[51] David Engel, Luciano Spinello, Rudolph Triebel, Roland Siegwart, Heinrich H Bülthoff, and Cristóbal Curio. Medial features for superpixel segmentation. In *MVA*, pages 248–252. Citeseer, 2009.

[52] Janine T Erler and Rune Linding. Network-based drugs and biomarkers. *The Journal of pathology*, 220(2):290–296, 2010.

[53] Shi et al. The MicroArray Quality Control (MAQC) project shows inter- and intraplatform reproducibility of gene expression measurements. *Nat Biotechnol*, 24(9):1151–1161, September 2006.

[54] EuroBoNet. Eurobonet, 2015.

[55] Richard Frank and Richard Hargreaves. Clinical biomarkers in drug discovery and development. *Nat Rev Drug Discov*, 2(7):566–580, 2003.

[56] Terrence S Furey, Nello Cristianini, Nigel Duffy, David W Bednarski, Michel Schummer, and David Haussler. Support vector machine classification and validation of cancer tissue samples using microarray expression data. *Bioinformatics*, 16(10):906–914, 2000.

[57] Salvador Gabarda and Gabriel Cristóbal. Blind image quality assessment through anisotropy. *JOSA A*, 24(12):B42–B51, 2007.

[58] Robin Genuer, Jean-Michel Poggi, and Christine Tuleau-Malot. Variable selection using random forests. *Pattern Recognition Letters*, 31(14):2225–2236, 2010.

[59] F Luengo Gimeno, SC Gatto, JO Croxatto, JI Ferro, and JE Gallo. In vivo lamellar keratoplasty using platelet-rich plasma as a bioadhesive. *Eye*, 24(2):368–375, 2010.

[60] Estelle Glory and Robert F Murphy. Automated Subcellular Location Determination and High-Throughput Microscopy. *Developmental Cell*, 12(1):7–16, January 2007.

[61] Dan B Goldman. Vignette and exposure calibration and compensation. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 32(12):2276–2288, 2010.

[62] Laurent Guigues, Hervé Le Men, and Jean-Pierre Cocquerez. The hierarchy of the cocoons of a graph and its application to image segmentation. *Pattern Recognition Letters*, 24(8):1059–1066, 2003.

[63] Isabelle Guyon, Jason Weston, Stephen Barnhill, and Vladimir Vapnik. Gene selection for cancer classification using support vector machines. *Machine learning*, 46(1-3):389–422, 2002.

[64] Douglas Hanahan and Robert A Weinberg. The hallmarks of cancer. *cell*, 100(1):57–70, 2000.

[65] Douglas Hanahan and Robert A Weinberg. Hallmarks of cancer: the next generation. *Cell*, 144(5):646–674, 2011.

[66] David Hasler and Sabine E Susstrunk. Color handling in panoramic photography. In *Photonics West 2001-Electronic Imaging*, pages 62–72. International Society for Optics and Photonics, 2000.

[67] Saima Hassan, Cristiano Ferrario, Aline Mamo, and Mark Basik. Tissue microarrays: emerging standard for biomarker validation. *Curr Opin Biotechnol*, 19(1):19–25, February 2008.

[68] Kai He, Ping-Fan Tang, and Ran Liang. Vignetting image correction based on gaussian quadrics fitting. In *Natural Computation, 2009. ICNC'09. Fifth International Conference on*, volume 5, pages 158–161. IEEE, 2009.

[69] Geoffrey E Hinton and Ruslan R Salakhutdinov. Reducing the dimensionality of data with neural networks. *Science*, 313(5786):504–507, 2006.

[70] André Homeyer, Andrea Schenk, Janine Arlt, Uta Dahmen, Olaf Dirsch, and Horst K Hahn. Practical quantification of necrosis in histological whole-slide images. *Computerized Medical Imaging and Graphics*, 37(4):313–322, 2013.

[71] Jan Hosang, Rodrigo Benenson, and Bernt Schiele. How good are detection proposals, really? *arXiv preprint arXiv:1406.6962*, 2014.

[72] Chih-Wei Hsu and Chih-Jen Lin. A comparison of methods for multiclass support vector machines. *Neural Networks, IEEE Transactions on*, 13(2):415–425, 2002.

[73] Kai Huang and Robert F Murphy. Boosting accuracy of automated classification of fluorescence microscope images for location proteomics. *Bmc Bioinformatics*, 5(1):78, 2004.

[74] Won-Ki Huh, James V Falvo, Luke C Gerke, Adam S Carroll, Russell W Howson, Jonathan S Weissman, and Erin K O'Shea. Global analysis of protein localization in budding yeast. *Nature*, 425(6959):686–691, 2003.

[75] Aapo Hyvärinen, Juha Karhunen, and Erkki Oja. *Independent component analysis*, volume 46. John Wiley & Sons, 2004.

[76] Anil Jain and Douglas Zongker. Feature selection: Evaluation, application, and small sample performance. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 19(2):153–158, 1997.

[77] Canny John. A Computational Approach to Edge Detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-8(6):679–698, 1989.

[78] Stephen C Johnson. Hierarchical clustering schemes. *Psychometrika*, 32(3):241–254, 1967.

[79] Ian Jolliffe. *Principal component analysis*. Wiley Online Library, 2002.

[80] Chris Jones, Alan Mackay, Anita Grigoriadis, Antonio Cossu, Jorge S Reis-Filho, Laura Fulford, Tim Dexter, Susan Davies, Karen Bulmer, Emily Ford, et al. Expression profiling of purified normal human luminal and myoepithelial breast cells identification of novel prognostic markers for breast cancer. *Cancer research*, 64(9):3037–3045, 2004.

[81] Thouis R Jones, Anne Carpenter, and Polina Golland. Voronoi-based segmentation of cells on image manifolds. In *Computer Vision for Biomedical Image Applications*, pages 535–543. Springer, 2005.

[82] Dejan Karadaglić and Tony Wilson. Image formation in structured illumination wide-field fluorescence microscopy. *Micron*, 39(7):808–818, 2008.

[83] Imen Karoui, Ronan Fablet, J-M Boucher, and J-M Augustin. Region-based image segmentation using texture statistics and level-set methods. In *Acoustics, Speech and Signal Processing, 2006. ICASSP 2006 Proceedings. 2006 IEEE International Conference on*, volume 2, pages II–II. IEEE, 2006.

[84] Peter M Kasson, Johannes B Huppa, Mark M Davis, and Axel T Brunger. A hybrid machine-learning approach for segmentation of protein localization data. *Bioinformatics*, 21(19):3778–3786, 2005.

[85] Ron Kohavi and George H John. Wrappers for feature subset selection. *Artificial intelligence*, 97(1):273–324, 1997.

[86] Daphne Koller and Nir Friedman. *Probabilistic graphical models: principles and techniques*. MIT press, 2009.

[87] Philipp Krähenbühl and Vladlen Koltun. Geodesic object proposals. In *Computer Vision–ECCV 2014*, pages 725–739. Springer, 2014.

[88] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

[89] Aparna Kumar, Arvind Rao, Santosh Bhavani, Justin Y Newberg, and Robert F Murphy. Automated analysis of immunohistochemistry images identifies candidate location biomarkers for cancers. *Proceedings of the National Academy of Sciences*, 111(51):18249–18254, 2014.

[90] Shawn Lankton, Delphine Nain, Anthony Yezzi, and Allen Tannenbaum. Hybrid geodesic region-based curve evolutions for image segmentation. In *Medical Imaging*, pages 65104U–65104U. International Society for Optics and Photonics, 2007.

[91] Honglak Lee, Roger Grosse, Rajesh Ranganath, and Andrew Y Ng. Convolutional deep belief networks for scalable unsupervised learning of hierarchical representations. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pages 609–616. ACM, 2009.

[92] Marylène Lejeune, Joaquín Jaén, Lluis Pons, Carlos López, Maria-Teresa Salvadó, Ramón Bosch, Marcial García, Patricia Escrivà, Jordi Baucells, Xavier Cugat, et al. Quantification of diverse subcellular immunohistochemical markers with clinicobiological relevancies: validation of a new computer-assisted image analysis procedure. *Journal of anatomy*, 212(6):868–878, 2008.

[93] F J W-M Leong, M Brady, and J O'D McGee. Correction of uneven illumination (vignetting) in digital microscopy images. *Journal of clinical pathology*, 56(8):619–621, 2003.

[94] Ilya Levner and Hong Zhang. Classification-driven watershed segmentation. *Image Processing, IEEE Transactions on*, 16(5):1437–1445, 2007.

[95] Kang Li, Eric D Miller, Mei Chen, Takeo Kanade, Lee E Weiss, and Phil G Campbell. Cell population tracking and lineage construction with spatiotemporal context. *Medical image analysis*, 12(5):546–566, 2008.

[96] Xiaomin Li, Yuanyuan Wang, Yinhui Deng, and Jinhua Yu. Cell segmentation using ellipse curve segmentation and classification. In *Information Science and Engineering (ICISE), 2009 1st International Conference on*, pages 1187–1190. IEEE, 2009.

[97] Andy Liaw and Matthew Wiener. Classification and regression by randomforest. *R news*, 2(3):18–22, 2002.

[98] Gang Lin, Umesh Adiga, Kathy Olson, John F Guzowski, Carol A Barnes, and Badrinath Roysam. A hybrid 3d watershed algorithm incorporating gradient

cues and object models for automatic segmentation of nuclei in confocal image stacks. *Cytometry Part A*, 56(1):23–36, 2003.

[99] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3431–3440, 2015.

[100] Lit-Hsin Loo, Lani F Wu, and Steven J Altschuler. Image-based multivariate profiling of drug responses from single cells. *Nature methods*, 4(5):445–453, 2007.

[101] Carrie J Lovitt, Todd B Shelper, and Vicky M Avery. Advanced cell culture techniques for cancer drug discovery. *Biology*, 3(2):345–367, 2014.

[102] Ricardo Macarron, Martyn N Banks, Dejan Bojanic, David J Burns, Dragan A Cirovic, Tina Garyantes, Darren V S Green, Robert P Hertzberg, William P Janzen, Jeff W Paslay, Ulrich Schopfer, and G Sitta Sittampalam. Impact of high-throughput screening in biomedical research. *Nat Rev Drug Discov*, 10(3):188–195, March 2011.

[103] Joanne Markham and José-Angel Conchello. Parametric blind deconvolution: a robust method for the simultaneous estimation of image and blur. *JOSA A*, 16(10):2377–2391, 1999.

[104] Andriy Marusyk, Vanessa Almendro, and Kornelia Polyak. Intra-tumour heterogeneity: a looking glass for cancer? *Nature Reviews Cancer*, 12(5):323–334, 2012.

[105] Jiri Matas, Ondrej Chum, Martin Urban, and Tomás Pajdla. Robust wide-baseline stereo from maximally stable extremal regions. *Image and vision computing*, 22(10):761–767, 2004.

[106] James G McNally, Tatiana Karpova, John Cooper, and José Angel Conchello. Three-dimensional imaging by deconvolution microscopy. *Methods*, 19(3):373–385, 1999.

[107] Corbin E Meacham and Sean J Morrison. Tumour heterogeneity and cancer cell plasticity. *Nature*, 501(7467):328–337, 2013.

[108] Erik Meijering. Cell segmentation: 50 years down the road [life sciences]. *Signal Processing Magazine, IEEE*, 29(5):140–145, 2012.

[109] Ezgi Mercan, Selim Aksoy, Linda G Shapiro, Donald L Weaver, Tad Brunye, and Joann G Elmore. Localization of diagnostically relevant regions of interest in whole slide images. In *2014 22nd International Conference on Pattern Recognition (ICPR)*, pages 1179–1184. IEEE, 2014.

[110] Fernand Meyer. Topographic distance and watershed lines. *Signal processing*, 38(1):113–125, 1994.

[111] J Monaco, J Tomaszewski, M Feldman, Mehdi Moradi, Parvin Mousavi, Alexander Boag, Chris Davidson, Purang Abolmaesumi, and Anant Madabhushi. Detection of prostate cancer from whole-mount histology images using markov random fields. In *Workshop on Microscopic Image Analysis with Applications in Biology (in conjunction with MICCAI)*, 2008.

[112] Dipti Prasad Mukherjee, Nilanjan Ray, and Scott T Acton. Level set analysis for leukocyte detection and tracking. *Image Processing, IEEE Transactions on*, 13(4):562–572, 2004.

[113] Kevin P Murphy. *Machine learning: a probabilistic perspective*. MIT press, 2012.

[114] Robert F Murphy. An active role for machine learning in drug development. *Nature chemical biology*, 7(6):327–330, 2011.

[115] Kaustav Nandy, Prabhakar R Gudla, Ryan Amundsen, Karen J Meaburn, Tom Misteli, and Stephen J Lockett. Automatic segmentation and supervised learning-based selection of nuclei in cancer tissue images. *Cytometry Part A*, 81A(9):743–754, July 2012.

[116] Sumit K Nath, Kannappan Palaniappan, and Filiz Bunyak. Cell segmentation using coupled level sets and graph-vertex coloring. *Med Image Comput Comput Assist Interv*, 9(Pt 1):101–108, 2006.

[117] Olympus. Micorscope resource center. [Accessed on June 2015. http://www.olympusmicro.com/index.html], 2015.

[118] C Ortiz de Solorzano, E Garcia Rodriguez, A Jones, D Pinkel, J W Gray, D Sudar, and S J Lockett. Segmentation of confocal microscope images of cell nuclei in thick tissue sections. *Journal of microscopy*, 193(3):212–226, 1999.

[119] C Ortiz de Solorzano, R Malladi, SA Lelievre, and SJ Lockett. Segmentation of nuclei and cells using membrane related protein markers. *journal of Microscopy*, 201(3):404–415, 2001.

[120] Dirk Padfield, Jens Rittscher, Nick Thomas, and Badrinath Roysam. Spatio-temporal cell cycle phase analysis using level sets and fast marching methods. *Medical image analysis*, 13(1):143–155, 2009.

[121] Dirk Padfield, Jens Rittscher, Nick Thomas, Badrinath Roysam, and others. Spatio-temporal cell cycle phase analysis using level sets and fast marching methods. *Med Image Anal*, 13(1):143, 2009.

[122] James Pawley. *Handbook of biological confocal microscopy*. Springer, 2010.

[123] Tingying Peng, Lichao Wang, Christine Bayer, Sailesh Conjeti, Maximilian Baust, and Nassir Navab. Shading correction for whole slide image using low rank and sparse decomposition. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2014*, pages 33–40. Springer, 2014.

[124] Rainer Pepperkok and Jan Ellenberg. High-throughput fluorescence microscopy for systems biology. *Nature Reviews Molecular Cell Biology*, 7(9):690–696, 2006.

[125] Zachary E Perlman, Timothy J Mitchison, and Thomas U Mayer. High-content screening and profiling of drug activity in an automated centrosome-duplication assay. *Chembiochem*, 6(1):145–151, 2005.

[126] Zachary E Perlman, Michael D Slack, Yan Feng, Timothy J Mitchison, Lani F Wu, and Steven J Altschuler. Multidimensional drug profiling by automated microscopy. *Science*, 306(5699):1194–1198, 2004.

[127] Zachary E Perlman, Michael D Slack, Yan Feng, Timothy J Mitchison, Lani F Wu, and Steven J Altschuler. Multidimensional drug profiling by automated microscopy. *Science*, 306(5699):1194–1198, 2004.

[128] F Piccinini, E Lucarelli, A Gherardi, and A Bevilacqua. Multi-image based method to correct vignetting effect in light microscopy images. *Journal of microscopy*, 248(1):6–22, 2012.

[129] Jordi Pont-Tuset, Pablo Arbelaez, Jonathan T Barron, Ferran Marques, and Jitendra Malik. Multiscale combinatorial grouping for image segmentation and object proposal generation. *arXiv preprint arXiv:1503.00848*, 2015.

[130] Nilanjan Ray and Scott T Acton. Data acceptance for automated leukocyte tracking through segmentation of spatiotemporal images. *Biomedical Engineering, IEEE Transactions on*, 52(10):1702–1712, 2005.

[131] Raghu Machiraju Rittscher, Jens and Stephen TC Wong. An introduction to fluorescence microscopy: Basic principles, challenges, and opportunities. In J. Rittscher, R. Machiraju, and S.T.C. Wong, editors, *Microscopic image analysis for life science applications*, Bioinformatics & Biomedical Imaging, chapter 4, pages 85–114. Artech House, Boston MA, USA, 2008.

[132] et al. Rizzardi, Anthony E. Quantitative comparison of immunohistochemical staining measured by digital image analysis versus pathologist visual scoring. *Diagn Pathol*, 2012.

[133] Nirupam Sarkar and BB Chaudhuri. An efficient differential box-counting approach to compute fractal dimension of image. *Systems, Man and Cybernetics, IEEE Transactions on*, 24(1):115–120, 1994.

[134] Dominik W Schelshorn, Armin Schneider, Wolfgang Kuschinsky, Daniela Weber, Carola Krüger, Tanjew Dittgen, Heinrich F Bürgers, Fatemeh Sabouri, Nikolaus Gassler, Alfred Bach, et al. Expression of hemoglobin in rodent neurons. *Journal of Cerebral Blood Flow & Metabolism*, 29(3):585–595, 2009.

[135] Bernhard Scholkopft and Klaus-Robert Mullert. Fisher discriminant analysis with kernels. *Neural networks for signal processing IX*, 1(1):1, 1999.

[136] Robert D Schreiber, Lloyd J Old, and Mark J Smyth. Cancer immunoediting: integrating immunitys roles in cancer suppression and promotion. *Science*, 331(6024):1565–1570, 2011.

[137] Aabid Shariff, Joshua Kangas, Luis Pedro Coelho, Shannon Quinn, and Robert F Murphy. Automated image analysis for high-content screening and analysis. *Journal of biomolecular screening*, 15(7):726–734, 2010.

[138] Reuben J Shaw and Lewis C Cantley. Ras, PI (3) K and mTOR signalling controls tumour cell growth. *Nature*, 441(7092):424–430, 2006.

[139] Tao Shi, David Seligson, Arie S Belldegrun, Aarno Palotie, and Steve Horvath. Tumor classification by tissue microarray profiling: random forest clustering applied to renal cell carcinoma. *Modern Pathology*, 18(4):547–557, 2005.

[140] Patrice Y Simard, Dave Steinkraus, and John C Platt. Best practices for convolutional neural networks applied to visual document analysis. In *null*, page 958. IEEE, 2003.

[141] Kevin Smith, Yunpeng Li, Filippo Piccinini, Gabor Csucs, Csaba Balazs, Alessandro Bevilacqua, and Peter Horvath. Cidre: an illumination-correction method for optical microscopy. *Nature methods*, 12(5):404–406, 2015.

[142] Malcolm A Smith, Nita L Seibel, Sean F Altekruse, Lynn AG Ries, Danielle L Melbert, Maura O'Leary, Franklin O Smith, and Gregory H Reaman. Outcomes for children and adolescents with cancer: challenges for the twenty-first century. *Journal of Clinical Oncology*, 28(15):2625–2634, 2010.

[143] Andrea Sottoriva, Inmaculada Spiteri, Sara GM Piccirillo, Anestis Touloumis, V Peter Collins, John C Marioni, Christina Curtis, Colin Watts, and Simon Tavaré. Intratumor heterogeneity in human glioblastoma reflects cancer evolutionary dynamics. *Proceedings of the National Academy of Sciences*, 110(10):4009–4014, 2013.

[144] Ernst HK Stelzer. The intermediate optical system of laser-scanning confocal microscopes. In *Handbook of biological confocal microscopy*, pages 207–220. Springer, 2006.

[145] Mammalian Gene Collection (MGC) Program Team et al. Generation and initial analysis of more than 15,000 full-length human and mouse cdna sequences. *Proceedings of the National Academy of Sciences*, 99(26):16899–16903, 2002.

[146] Laura Toloşi and Thomas Lengauer. Classification with correlated features: unreliability of feature ranking and solutions. *Bioinformatics*, 27(14):1986–1994, 2011.

[147] Elizabeth C Toomey, Joshua D Schiffman, and Stephen L Lessnick. Recent advances in the molecular pathogenesis of ewing's sarcoma. *Oncogene*, 29(32):4504–4516, 2010.

[148] M Von Tiedemann, Anders Fridberger, M Ulfendahl, and J Boutet De Monvel. Image adaptive point-spread function estimation and deconvolution for in vivo confocal microscopy. *Microscopy research and technique*, 69(1):10–20, 2006.

[149] Micah Webster, Keren L Witkin, and Orna Cohen-Fix. Sizing up the nucleus: nuclear shape, size and nuclear-envelope assembly. *Journal of cell science*, 122(10):1477–1486, 2009.

[150] Ran Xie, Joon-Yong Chung, Kris Ylaya, Reginald L Williams, Natalie Guerrero, Nathan Nakatsuka, Cortessia Badie, and Stephen M Hewitt. Factors Influencing the Degradation of Archival Formalin-Fixed Paraffin-Embedded Tissue Sections. *J Histochem Cytochem*, 59(4):356–365, April 2011.

[151] Lin Yang, Yizhe Zhang, Ian H Guldner, Siyuan Zhang, and Danny Z Chen. Fast background removal in 3d fluorescence microscopy images using one-class learning. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015*, pages 292–299. Springer, 2015.

[152] Xiaodong Yang, Houqiang Li, and Xiaobo Zhou. Nuclei segmentation using marker-controlled watershed, tracking using mean-shift, and kalman filter in time-lapse microscopy. *Circuits and Systems I: Regular Papers, IEEE Transactions on*, 53(11):2405–2414, 2006.

[153] Anas Younes and Donald A Berry. From drug discovery to biomarker-driven clinical trials in lymphoma. *Nature Reviews Clinical Oncology*, 9(11):643–653, 2012.

[154] Wonpil Yu, Yunkoo Chung, and Jung Soh. Vignetting distortion correction method for high quality digital imaging. In *Pattern Recognition, 2004. ICPR 2004. Proceedings of the 17th International Conference on*, volume 3, pages 666–669. IEEE, 2004.

[155] C Zhang, C Sun, and T D Pham. Clustered nuclei splitting using curvature information. *Digital Image Computing ...*, 2011.

[156] Wei Zhang, Ilya Shmulevich, and Jaakko Astola. *Microarray Quality Control*. John Wiley & Sons, March 2005.

[157] C Lawrence Zitnick and Piotr Dollár. Edge boxes: Locating object proposals from edges. In *Computer Vision–ECCV 2014*, pages 391–405. Springer, 2014.

[158] JM Zwier, GJ Van Rooij, JW Hofstraat, and GJ Brakenhoff. Image calibration in fluorescence microscopy. *Journal of Microscopy*, 216(1):15–24, 2004.