CHAPTER 1

Introduction

1.1 Overview of the Research Work

Cytopathology is the study and diagnosis of diseases at the cellular level. The microscopic examination remains as the gold standard for cell analysis due to its low cost and widespread acceptance. However, the manual examination is a laborious task involving both slide preparation (fixation and staining) and analysis. It is a time consuming, repetitive and tedious job. Above all, the results may vary for the same sample among the clinicians depending on their level of expertise. In order to overcome these drawbacks, several efforts have been made in the recent past to automate the process of cytopathology. Research efforts in this direction have mostly been constrained to two approaches: automation of slide preparation and automation of slide analysis. Instruments which can carry-out automated slide preparation employ extensive amount of robotic handling, rendering the commercially available whole slide scanning and analysing systems bulky and expensive (Rojo et al. (2006); Pantanowitz et al. (2011)). One of the cheaply available automated microscopy systems is PathScope (PathScope (2016)) but costs around 25000 US $ making them not that affordable in resource limited clinics especially in low-income group countries.

The automated microscopy systems such as the PathScope (PathScope (2016)) has considerably improved the throughput (number of cells that can be processed in unit time) when compared to manual microscopy but is still well behind flow cytometry. Flow cytometry has become an indispensable tool for clinicians and biologists for counting, analysing and identifying cells with typical throughput of the order of a few thousands cells per second. It uses a flow cell architecture where the cells are interrogated using lasers while they are in flow. A typical flow cytometry system measures the forward as well as side scatter profiles of the lasers. The forward scatter is a measure of the size of the cell and the side scatter is a

measure of the complexity of the cell (refer Fig. 1.1), and this knowledge is used to identify and count different cells under study.

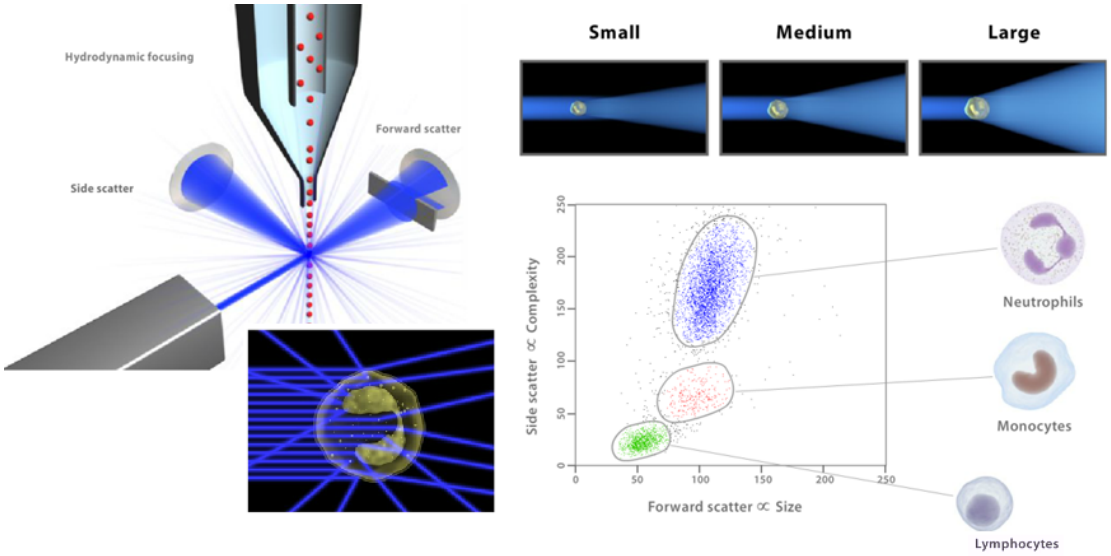

Figure 1.1: Analysis of white blood cells by flow cytometry (Flow (2016))

While the acquisition speed of flow cytometry is extremely high, the amount of information it provides per cell is usually low. The reason is that the flow cytometry will not capture specific morphological features other than the amount of scatter. On the other hand, traditional microscopic examination offers detailed images with spatial localisation of sub-cellular components but has drawbacks in terms of enumeration and speed. Imaging Flow Cytometry (IFC) (Basiji et al. (2007); Schonbrun et al. (2012)) is a nascent technology that combines the speed of flow cytometry and the power of digital microscopy in providing the capability to analyse morphological features. However, the current commercially available IFC systems are bulky and expensive. For example, the Amnis IFC by Merck Millipore Inc (Amn (2016)) costs around 199,000 US $. These systems use bulk fluid handling mechanisms for automating the process of sample image acquisition and employ sophisticated and expensive image acquisition schemes which use time-delay integration detectors and multiple laser sources. The machines are too expensive to afford in resource limited settings. The recent trend has been to employ microfluidic sample handling in combination with different microscopy imaging modalities to enable high-throughput imaging of cells in flow. These systems, which we refer as microfluidics microscopy (Mf-Ms), combine the statistical power of flow cytometry with spatial and quantitative morphology of digital

microscopy.

In this research, we propose processing frameworks for the data acquired from cost-effective whole slide analysing system and a reasonably high-throughput prototype microfluidics microscopy system (Jagannadh et al. (2016)) developed by our collaborators. Unlike the commercially available systems, the developed systems employ inexpensive off-the-shelf optical components and fluid handling mechanisms. This has enabled us to set-up portable, automated disease diagnostic/screening platforms for resource limited settings. Thus my research thesis has dual objectives. i) Analysis of stained blood smear images/videos captured using custom built automated full slide scanner, thereby providing a cost-effective solution for malaria diagnosis in conventional gold-standard microscopy. ii) Analysis and classification of cells from unstained IFC data. The common goal is to design and develop necessary framework containing advanced image analysis and machine learning techniques to operate on the data from very cost-effective instruments thereby moving in a direction to make portable diagnostic/screening systems for resource limited settings.

In this thesis, first we put forward necessary image analysis and classification algorithms for the custom-built automated whole slide scanner for detecting and quantifying textitPlasmodium falciparum infected malarial cases. This is accomplished by cell localisation by a proposed cascaded segmentation strategy and parasite detection using a custom-designed Convolutional Neural Network (CNN) on focus stack of slide images. Subsequent chapters discuss about the data analysis from a prototype Mf-Ms system. As noted, the Mf-Ms systems employ inexpensive optics, as well as polymer/plastic based microfluidic devices (around 1 $). Thus the total cost of the components of the microfluidics IFC system developed (Jagannadh et al. (2016)) is only about 1500 $ when compared to 199,000 $, the cost of Amnis (Amn (2016)) IFC. I have proposed, as part of this research, a general framework capable enough to automatically analyse the cells captured using custom-designed microfluidics microscopy systems and have also built a prototype for signature based as well as hand engineered feature based classification of unstained unlabeled Leukaemia cell-lines K562, MOLT, and HL60.

In the last chapters of this research report, we propose the use of deep learn-

ing based cell classifiers for better accuracy. The proposed framework is capable enough to deal with limited availability of labeled data for building a supervised classification system. The Restricted Boltzmann Machine (RBM) based deep belief network makes use of all available data (both labeled as well as unlabeled) to learn the underlying structure of the training data so that the subsequent supervised training needs only very few training samples for learning the classifier. Also, we propose to use the transfer learning capability of CNN to extract sensible and discriminative features to produce better accuracy for the leukaemia cell-line classification.

1.2 Motivation and Scope

There is a great demand and huge medical value for cost-effective cytopathology. As discussed in last section, the available automated cytopathology systems are bulky and unaffordable to many clinics especially in poor-income group countries. This has motivated us to define the scope of our research and we set our goal to design and develop necessary image analysis and pattern recognition techniques for setting up low-cost, portable instruments for point-of-care diagnostics.

1.3 Contributions of the Thesis

The goal of this research work, as noted in section 1.2, is to propose necessary image analysis and pattern recognition techniques for low-cost automation for cytopathology. Towards this goal, we have made the following contributions through this research.

- A fully automated quantification system for the diagnosis of malaria due to protozoan of type *Plasmodium falciparum* from focus stack of blood smear images collected using a cost-effective, custom-built, portable whole slide scanner is proposed. A custom designed convolutional neural network is used to detect the malaria infected cells. Use of CNN operating on focus stack for the detection of malaria is first of its kind, and it not only improved the detection accuracy both in terms of sensitivity and specificity

but also favoured the processing on cell patches and avoided the need for hand-engineered features. The proposed approach (portable slide scanner and the CNN algorithm together) is suitable for point-of-care Diagnostics and eliminates the need for a pathologist to manually examine the slides using a bright-field microscope.

- The proof-of-concept of a diagnostic framework as well as a prototype for signature based analysis and classification of leukaemia cell lines (K562, MOLT, HL60) imaged using custom fabricated, cost-effective microfluidics imaging flow cytometry (mIFC) is proposed. The mIFC is an emerging technology that combines the statistical power of flow cytometry with spatial and quantitative morphology of digital microscopy. We have analysed the feasibility of a cell signature based approach for cancerous cell identification analogous to face recognition systems implemented with help of surveillance cameras. We have also proposed a way by which cells difficult to classify can be identified, which will open up an opportunity to clinicians to correctly identify the true class of the cell rather than going for a wrong classification. Altogether, such a platform would facilitate affordable mass screening camps in the developing countries and therefore help to decrease cancer mortality rate.

- We Proposed a general framework for the processing/classification of cells in mIFC. The framework includes computationally feasible cell preprocessing methods and a proposed graph based cell segmentation strategy to find the contour of the cell. Once the cells are localised, a set of features reflecting the size, shape, and complexity of the cells are extracted and are used to classify the cells. The usefulness of the framework is established by performing the classification of leukaemia cell-lines. The proposed system is a significant development in the direction of building cell analysis platform that would facilitate affordable mass screening camps looking cellular morphology for disease diagnosis.

- We have also explored and proposed the feasibility of using deep learning networks for cytopathology by performing the classification of leukaemia cell-lines. The capability of Restricted Boltzmann Machine (RBM) based

systems in learning the structure of the data rather than learning labels is utilized to build a Deep Belief Network (DBN) for classification. The transfer learning capability of deeply trained CNN on extensive non-medical image database such as ImageNet is made useful to successfully classify the leukaemia cell-lines K562, MOLT and HL60. These capabilities of both RBM and CNN are very useful in medical image domain where often large dataset is available for training but only a small fraction is labeled, limiting the capability of building supervised deep learning based classifiers. In our investigation, we have found that the proposed methods outperformed the conventional systems in the classification of these cell lines. To the best of our knowledge, such a reporting on cytopathology images is first of its kind and we believe that it holds great promise in terms of enabling cancer screening in resource-poor settings.

1.4 Organisation of the Thesis

Chapter 2 explore the cellular image diagnosis in general and microscopy as well as microfluidic microscopy based cytopathology in particular. Chapter 3 describes the image analysis and classification system proposed for automated quantitative malaria diagnosis, using a custom developed whole slide scanning system. Chapter 4 introduces the nascent imaging flow cytometry and covers the details of the prototype developed for signature based leukaemia cell-line analysis. Chapter 5 provides the general framework proposed for the analysis and subsequent classification of cells captured using the custom-built microfluidic microscopy systems. Chapter 6 proposes the use of deep learning systems for better classification accuracy. It provides necessary details for using RBM and CNN for the use in cytopathology to build deep learning based classification system even with limited labeled training data. Each chapter starts with the goal that it is trying to accomplish, and covers necessary literature, discusses the contributions. Chapter 7 summarizes the thesis and suggests directions for future works.