University of Alberta

Data based abnormality detection

by

Yashasvi Purwar

A thesis submitted to the Faculty of Graduate Studies and Research

in partial fulfillment of the requirements for the degree of

Master of Science

in

Process Control

Department of Chemical and Materials Engineering

Spring 2011

Edmonton, Alberta

Examining Committee

Prof. Sirish L. Shah
Chemical and Materials Engineering, University of Alberta

Prof. Stevan Dubljevic
Chemical and Materials Engineering, University of Alberta

Prof. Nilanjan Ray
Department of Computing Science, University of Alberta

Prof. Amos Ben-Zvi
Chemical and Materials Engineering, University of Alberta

Dedicated to

Mummy, Daddy, Kirti and Sudhakar, your love, support and understanding has been my source of strength

Acknowledgements

Guru Brahma Gurur Vishnu
Guru Devo Maheshwaraha
Guru Saakshat Para Brahma
Tasmai Sree Gurave Namaha

I am heartily thankful to my supervisor Dr. Sirish L Shah, for his guidance and above all the freedom I enjoyed under his supervision. I am overwhelmed with gratitude for the motivation and support he provided throughout the course of study. The freedom allowed me to explore contemporary areas of research in detail. The experience gained in exploring various research fields has boosted my academic growth. It helped me understand my strengths and gave confidence to face life with all its joys and sorrows.

I would also like to thank Dr. Gwen Clarke and Dr. Areej Almugairi for their support in understanding the field of microscopy. The extensive discussions with them motivated me to come up with a solution for automatic diagnosis. From numerous discussions I have gained beyond what is presented in this thesis. This thesis would not have been possible without the support of Dr. Phanindra Varma and Karteek Popuri. The discussions with Dr. Phanindra helped me to gain substantial amount of image processing knowledge and significant direction for my research. I am indebted to friends because of whom stay at Edmonton became an unforgetTable memory. I would like to thank Adi, Amit, Arjun, Anuj, Bhusie, Dipen, Dru, Gujju, Johnny, Mama, Maede,Manjeet, Nagi, Phanindra, Rishi, Samidh, Saneej, Sandy, Sandeep, Shishir and Tunil.

Abstract

Data based abnormality detection is a growing research field focussed on extracting information from feature rich data. They are considered to be non-intrusive and non-destructive in nature which gives them a clear advantage over conventional methods. In this study, we explore different streams of data based anomalies detection. We propose extension and revisions to existing valve stiction detection algorithm supported with industrial case study. We also explored the area of image analysis and proposed a complete solution for Malaria diagnosis. The proposed method is tested over images provided by pathology laboratory at Alberta Health Service. We also address the robustness and practicality of the solution proposed.

Table of contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Motivation

'Abnormality' literally is defined as behaviour that breaches the rule. As the definition suggests the behaviour of systems (human body can also be considered as a system) or processes operating within these rules are acceptable. Thus it becomes important to monitor a system to avoid deviation or abnormal behaviour to boost profitability. As we walk in the era of digital age or in the age of information technology, system monitoring has become excessively widespread. The system's information is delivered in the form of digital data like time series data of critical variables, or digital images etc. The aim of the study presented here is to extract abnormalities or anomalies present in the digital data which cause deviation from normal working conditions. The problems arising from the abnormal behaviour needs to be quantified to understand the severity of the problem to help in prioritising the maintenance of such systems. Thus, it becomes important to understand data more closely as it carries physical information of the system and develop smarter data driven techniques to capture the maximum amount of information about the system from the delivered data. Data driven techniques have a clear advantage in industry as well as medicine because they are non-intrusive in nature. The two case studies presented in this research work show the need for data based diagnosis methods for detection of anomalies.

1.2 The valve problem

Typical industrial plants such as an oil refinery include thousands or more control loops to operate the process at desired operating conditions. The first case study presented in this thesis focuses on the performance of control loops. In particular, a poorly performing control loop results in degraded process performance. The degraded process performance results in poor plant performance which directly impacts the profitability of the plant. Thus control loop monitoring has emerged

as an important capital asset to industry. Most control loops are equipped with movable mechanical equipment, the actuator, to control the process and the most commonly used actuating device is the control valve. Many literature studies suggest that 20-30% [1 and 8] of all control loops have oscillation problems which can directly be related to valve problems. The physical understanding of the valve indicates that *st*atic fr*iction* (stiction) is most often the cause for poor control loop performance. A typical valve controlled loop is shown in Figure 1.1. The stiction induced non-linear oscillations known as limit cycles [1 and 8] are shown in Figure 1.2 leading to system instability.

Choudhury *et al* [1] introduced a bi-coherence based data driven method to detect and quantify stiction and this method is reported to be most successful method for stiction detection [8]. However, the bi-coherence method often gives a number of false and missed alarms and the primary focus of their method is stiction detection and quantification. The aim of the current research in this thesis is to rename the 'Valve Stiction' detection algorithm as 'valve fault detection' algorithm and also minimize the number of missed and false alarms.

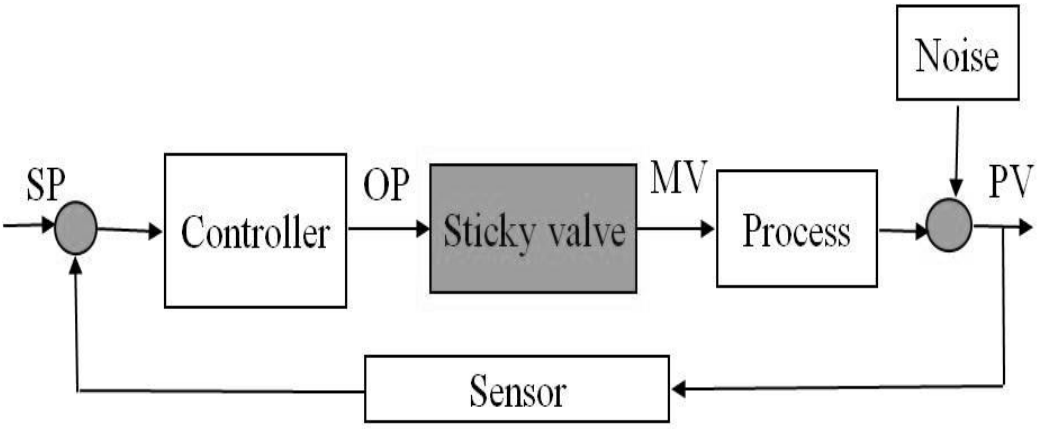


Figure 1.1: Schematic of simple feedback of valve controlled process

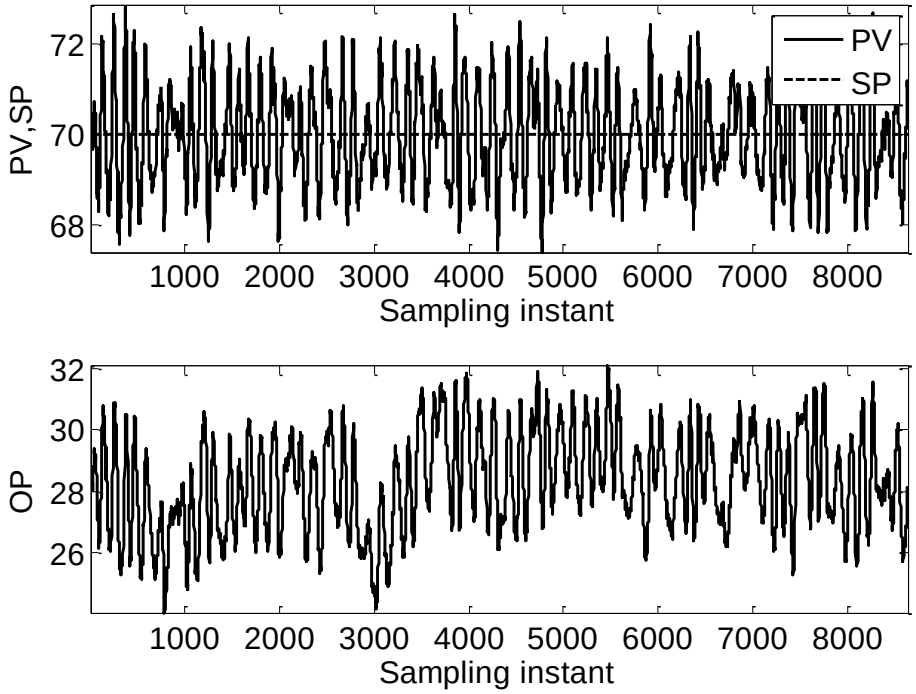
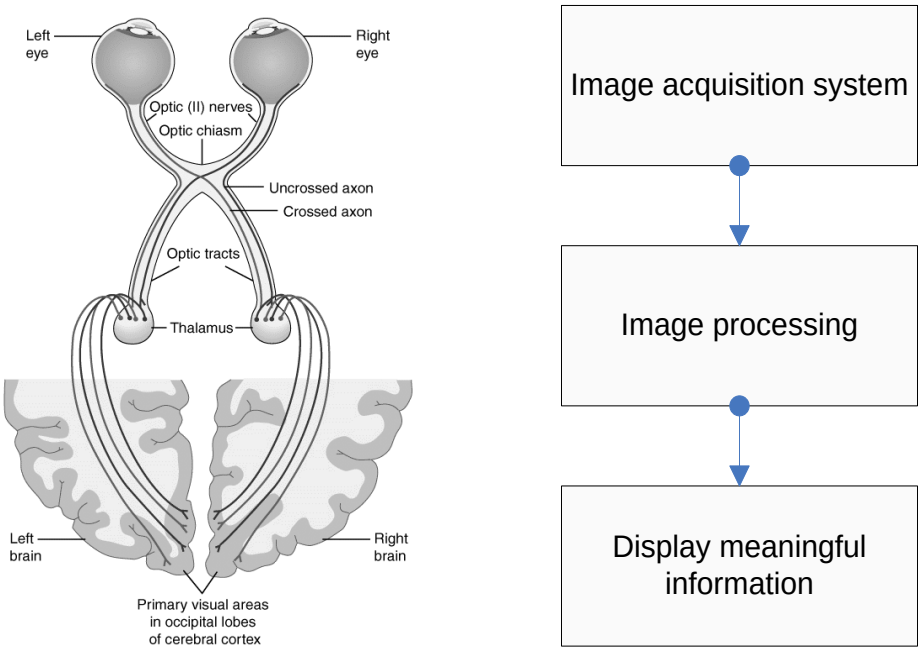


Figure 1.2: The limit cycles due to valve problems in PC950 loop from industrial data set

1.3 Targeted information capture in digital images

The second case study in this thesis primarily explores the field of image processing to provide computer aided vision methods to detect the presence of malaria parasites in thin smeared blood sample slides. Humans are gifted with an exceptionally powerful visual system which makes them experts in visual based judgement. Although the foundation of image processing is mathematics and probabilistic formulations, human intuition and analysis plays an important role in the choice of particular method over another. Hence, we first try to understand the basics of human vision system. Figure 1.3a shows a simple schematic of human visual system which comprises of the eyes, optic nerves and visual cortex at the rear of our brain. The eye captures visual data and transfers this to the visual cortex through the optic nerve. The brain interprets the visual stimuli which form a large part of our experience. Figure 1.3b show an overly simple imitation of the human visual system.



(a) (b)

Figure 1.3: Imitation of Human visual system with computer aided vision.

The emergence of low cost digital imaging system has given access to cheap and feature rich information for automatic image analysis. Image analysis has opened a huge field for automation, wherein algorithms are designed to extract information from digital images for subsequent decision support. Automatic image analysis is very diverse, popular and has deep roots in the medical field such as MRI scans for locating tumours. The focus of this work is to introduce the effectiveness of automated image analysis in the field of malaria diagnosis.

Malaria is the third deadly infectious disease [6] and has infected humans for almost 500,000 years [5]. High death tolls and large volume of infections especially in the developing countries as shown in Figure 1.4 clearly motivate the need to control the disease by providing cheap and speedy diagnosis methods combined with effective medication. The diagnosis of malaria is still in its primitive stage and extensively dependent on manual microscopy. The overall understanding of manual microscopy reveals a huge potential for image analysis

based automation methods. The semi-automatic diagnosis process is designed to imitate the process of manual microscopy as closely as possible. Also, the semi-automatic technique as developed here has specifically been designed under the constraint that it should be unsupervised, specific and practical.

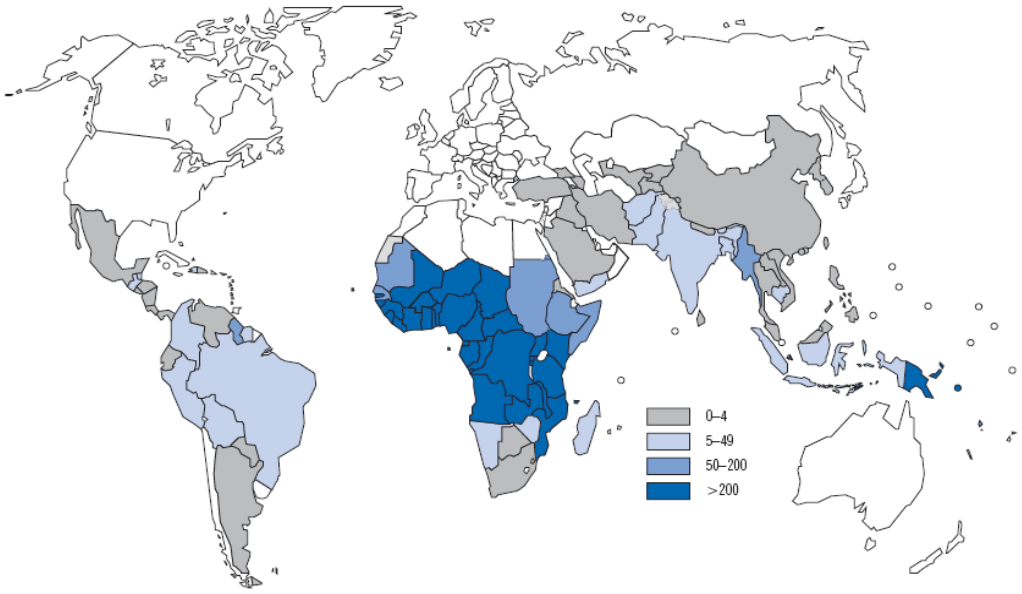


Figure 1.4: Estimated incidence of malaria per 1000 population, 2006 [5]

The method proposed in this thesis is based on image analysis combined with robust statistical methods to capture the specific information in the low resolution images. The method was tested on a large variety of images provided by personnel at Alberta Health Services.

1.4 Thesis overview

This thesis is organised as follows: chapter 2 discuses the basics of valve controlled loops and the problems related to valves. Specifically, this chapter discusses the bi-coherence based method proposed by *Choudhoury et al* [1] and proposes improvements to the current bi-coherence method. The importance of revisions and improvements suggested are tested through the two industrial case studies presented later in the chapter.

Chapter 3 starts by motivating the need for an automated method for malaria diagnosis. This chapter discusses the development of novel image based diagnosis method with detailed explanation of each step. The chapter carefully looks at the choice of method for each step under the given constraint. The success rate of the method is discussed in the later part of chapter by analysing more than 200 images received from the University of Alberta hospital.

The thesis ends with concluding remarks in Chapter 4 with a short discussion on suggestion for future work.

Chapter 2

Bi-coherence Based Valve Stiction Detection: Extensions and Recommendations

2.1 Introduction

The most common actuating element in a control loop is a valve which is typically the only moving element in a feedback loop. A cross-sectional diagram of a typical pneumatic control valve is shown in Figure 2.1. The valve represented in this Figure portrays the motion of a spring-diaphragm actuated stem in a flow control loop. The valve stem is surrounded by tight packing to ensure that fluid flow under control does not leak. Because of this there is static friction between the packing and the moving valve stem. Thus, valve stiction (as a short from for *st*atic fr*iction*, to distinguish from dynamic friction) is one of the most common causes of non-linearity in the control loop. Typically, sticky valves will cause limit cycles in the loop and these oscillations can easily propagate to other loops in an interconnected system. In addition to this there may be other problems with the actuator such as problems with the valve positioner, ruptured diaphragms, broken springs, hysteresis, backlash and possible non-linear excursions of the process. The objective of this study is to therefore be able to develop a general data-based 'valve fault detection' tool. The implicit assumption here is that the process can be considered to be locally linear and therefore any non-linearity detected in a loop can be attributed to valve problems.

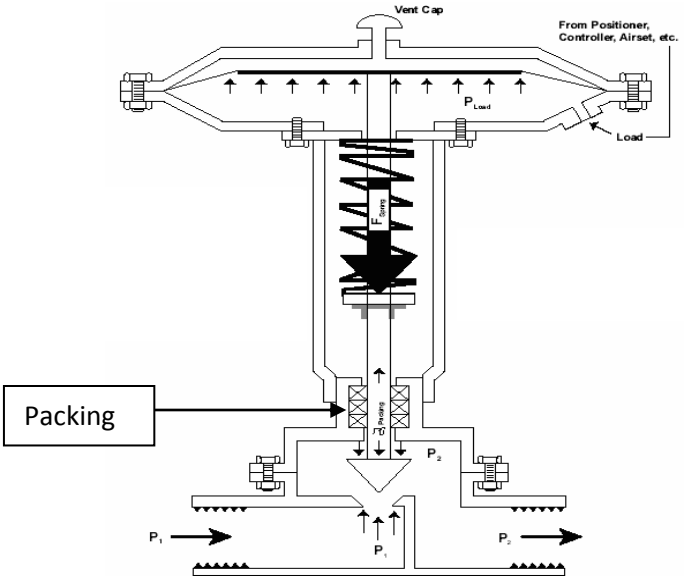


Figure 2.1: Spring Diaphragm valve

2.2 A closer look at stiction

Detection and quantification of valve problems is therefore an important issue in industry in order to prioritize the maintenance of control loop hardware to enhance plant performance. Among all of this, valve stiction is a fairly wide spread problem and as a result there have many studies in the literature suggesting several methods for detection and quantification of valve stiction [8]. Figures 2.2 and 2.3 give a graphical illustration of valve stiction as a fault detection and fault diagnosis problem. To detect and quantify stiction it is first necessary to understand the physical behavior of sticky valve. The physical behavior of valve stiction can be closely captured using a two parameter model shown in Figure 2.3, as proposed by Choudhury et al [1]. The two parameters are S (stick band + dead band) and J (slip jump) as shown in Figures 2.3 and 2.4.

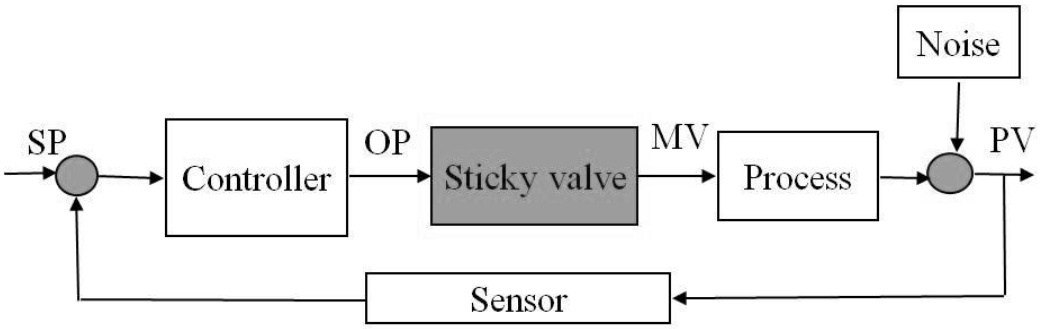


Figure 2.2: Control loop diagram with a sticky valve

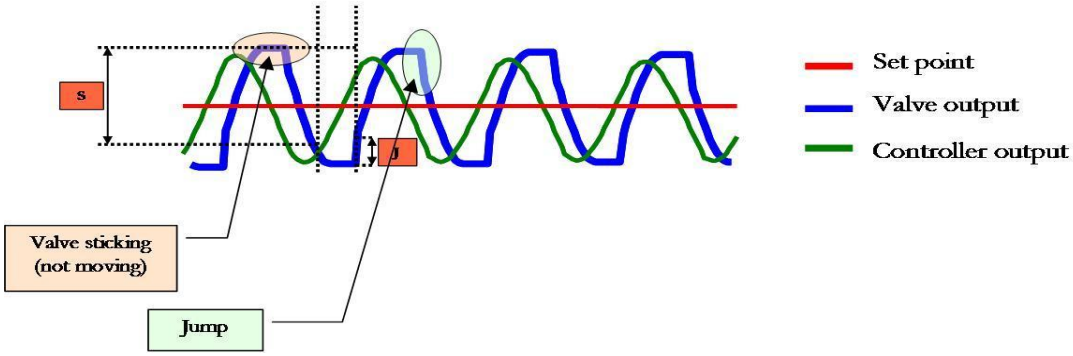


Figure 2.3: Sticky valve output showing the magnitude of stick band plus deadband (S) and slip-jump (J) parameters

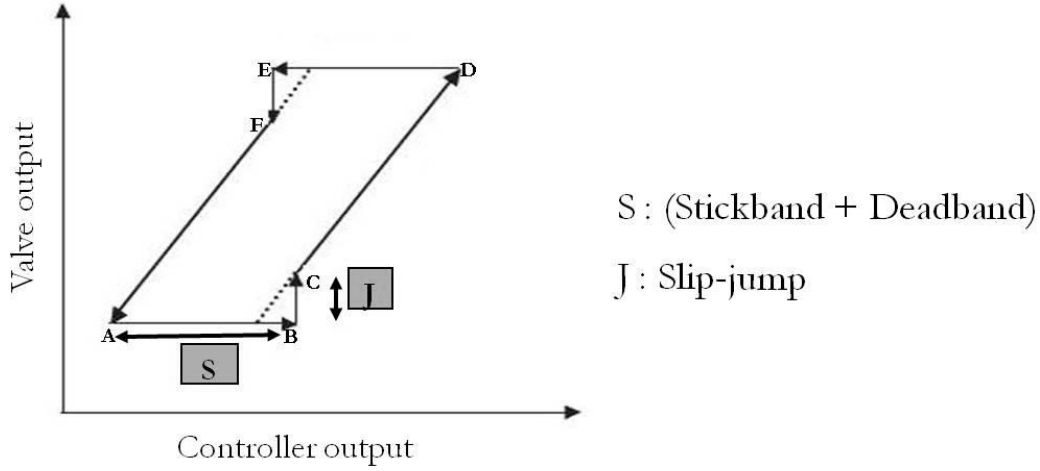


Figure 2.4: Typical input-output behavior of a sticky valve

2.3 Extensions and Improvements

The bi-coherence based non-intrusive stiction detection algorithm has been reported [8] as one of the most successful methods for detecting and quantifying stiction. The algorithm uses process output (PV), controller output (OP) and set point (SP) data to detect and quantify stiction. However the bi-coherence based algorithm in its current form [1 and 8] results in a number of false and missed alarms. The objective in this section is to propose extensions to the existing valve stiction algorithm to minimize the number of missed and false alarms and thus enhance the performance and robustness of the algorithm. The reasons for the false and missed alarms and the corresponding extensions to overcome the limitations of the current algorithm are listed below and are supported by appropriate examples to demonstrate the utility of the proposed changes on simulated and industrial data sets. Details of the suggested changes are also given in MATLAB code.

1) *Detrending of data:* The algorithm starts with the detection of oscillation in the error signal (SP-PV). Oscillation detection is very sensitive to the stationarity of the signals that are being analyzed; to make the oscillation detection robust, it is advisable to make the error signal stationary by detrending. Detrending, removes the mean value of time series and makes oscillation detection robust and less prone to errors, this reduces the number of false alarms. In all of the following examples, we use detrended data. We start with one simple example to illustrate one such case:

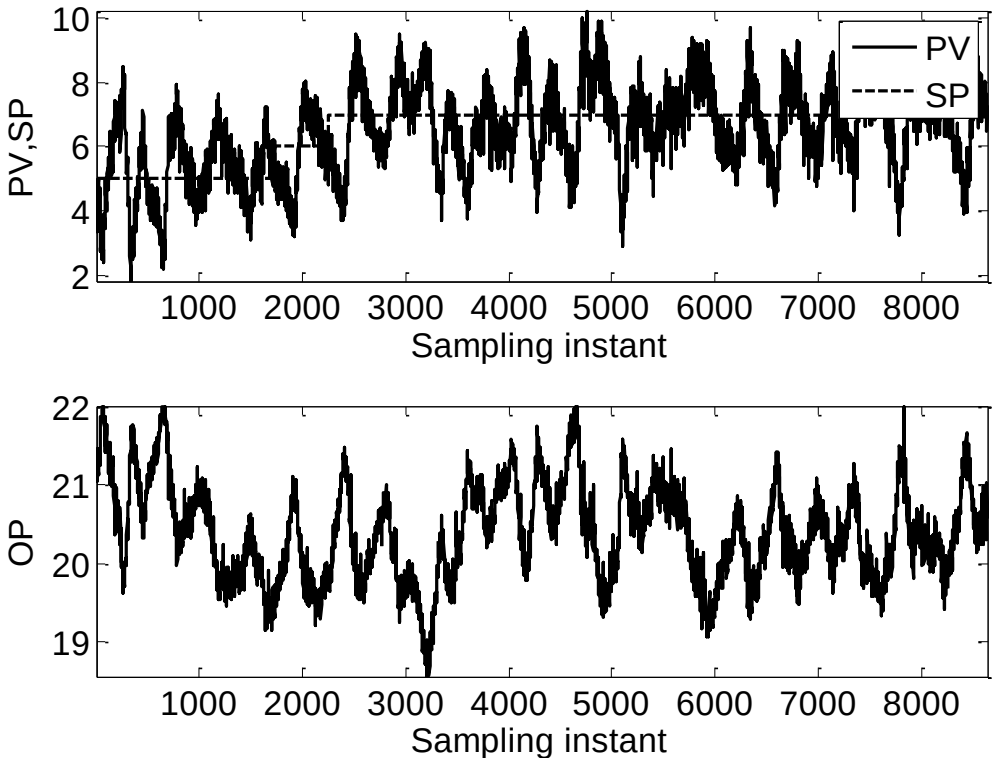


Figure 2.5: The original signal (for tag PC2345 from an industrial site*) was non-stationary and the stiction detected was positive. However there was no stiction detected in this valve after detrending; the algorithm thereafter correctly diagnoses this case as a 'no stiction' signal.

* Industrial site name withheld for confidentiality reasons.

2) *Non-Linearity check:* Non linearity detection is one of the initial checks in the stiction detection and quantification algorithm; it fails to detect and quantify stiction if the computed non-linearity index (NLI) of the error signal is non-positive. NLI is sensitive to the number of data points and changes if the data is segmented into smaller over-lapping windows particularly in the case of noisy data. Perhaps, the reason for failure of detecting nonlinear 'elements' in a large dataset, meaning many observations, is that the NLI is averaged. As a result the NLI may be lower over the entire data set when some windows segments may have linear characteristics and even though other window segments may indicate

significant nonlinearities. If on the other hand, NLI is computed on segments of data that are created by overlapping windows, each of length at least 1024 points, then the NLI of many of these windows may be larger. To make the NLI more robust in such cases, it is suggested that the data be segmented as follows with an agent based voting decision.

- For observations less than 1024 points, report it as insufficient data for quantification.
- For observations between 1024 to 2048 points, choose segments of width of 1024 points with moving overlapping width of 100 points. Than the final NLI decision of 'yes' or 'no' is arrived at by vote count.
- For observations with more than 2048 points, choose segments of width of 2048 points with moving overlapping width of 200 points. Than the final NLI decision of 'yes' or 'no' is arrived at by vote count.

This approach circumvents the requirement of the data being free from sharp or abrupt changes. For example, if the data comes from a loop with no valve or non-linearities but has some abrupt or sharp changes then the windows containing these changes will indicate high non-linearity. However a majority of the windows will not show the presence of any non-linearity. This feature clearly serves to extend the bi-coherence method and overcomes the problem that was commented in Ref [8] page number 323 *'requirement of a large dataset free from sharp or abrupt changes are clear weakness of the method'*.

In this way it is suggested that data segmentation for NLI and stiction quantification be differentiated i.e. NLI is calculated over large number of overlapping windows and if non-linearity is detected then stiction is quantified over the data segment where NLI is the highest. In order to make the diagnostics easier and also to extract maximum information from the analysis, the algorithm

is extended to **also** report percentage of segmented windows with high NLI as explained below:

- If more than half segmented windows are detected to have high NLI then maximum NLI is reported along with the percentage of segmented data with high NLI and stiction is quantified.
- If less than half segmented windows report high NLI then the algorithm will only report the percentage of segmented data with high NLI.

These changes were applied and checked on the industrial dataset shown below. In both cases, segments of width 2048 points with moving overlapping width of 200 points were selected:

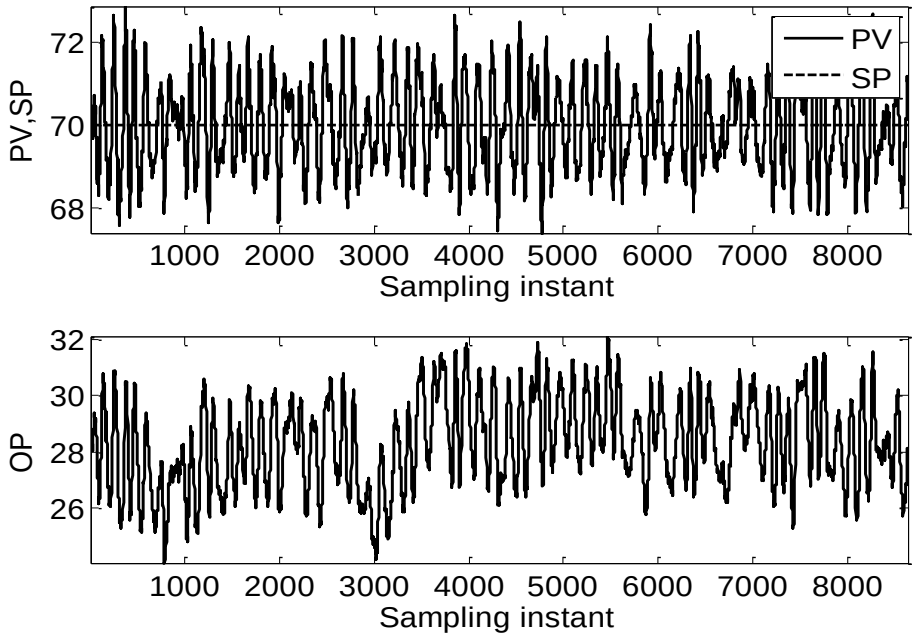


Figure 2.6 *(*Tag PC245): Initially without NLI calculations on segmented data no stiction was detected. As can be visualized there are limit cycles in this data because of valve problems. After implementing the suggested changes, the NLI is found to be 0.52 thus allowing the algorithm to confirm non

linearity and correct diagnosis of stiction. In this case percentage of window segments with high NLI is reported to be 58%.

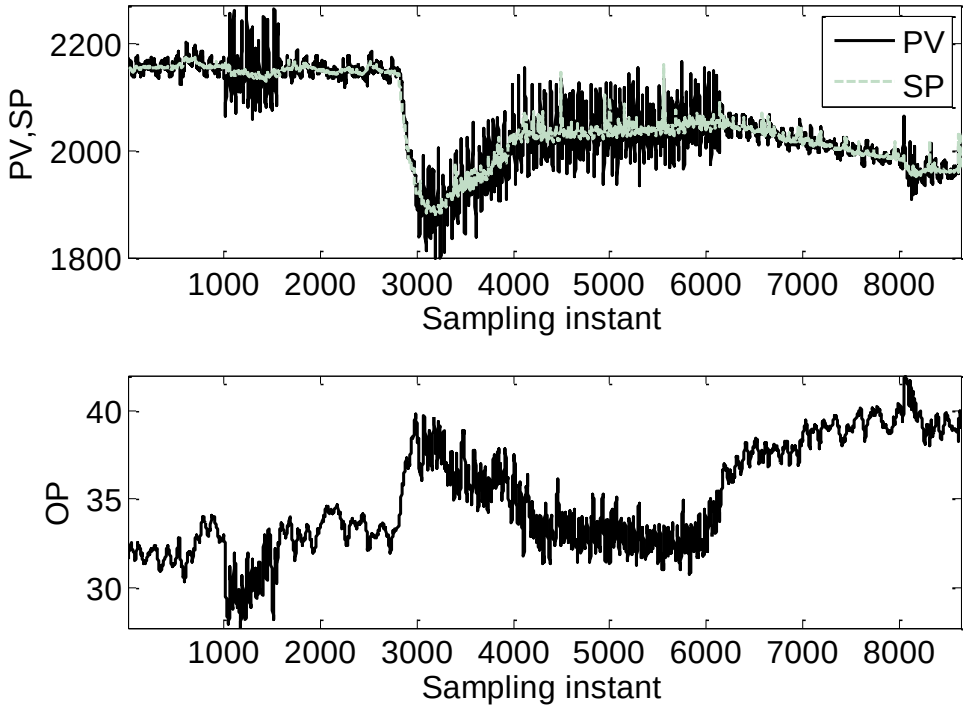


Figure 2.7 (Tag: FC933): The presence of limit cycles clearly shows valve problems in some data segments with maximum NLI value of 0.89 thus confirming stiction or valve problems or nonlinear process excursions. In this case percentage of window segments with high NLI is reported to be 64%

3) *Set point activity check:* Oscillations in the set point (SP) values can be one of the causes of process variable (PV) oscillations. This could potentially be caused by problems in the outer loop in a cascaded loop setting. The present algorithm has a check for set point activity but we suggest an extra step to make the SP activity check more robust. The current algorithm uses correlation between SP and PV blindly, not taking into account the time-lag between SP and PV. It is necessary to use lag-adjusted correlation to minimise the number of missed and false alarms.

Alternately, we suggest use of coherency information; the coherency matrix tries to capture the correlation between two time series in the frequency domain thus making it phase blind. Coherency analysis between the SP and PV will therefore be 'blind' to the time lag between the two time series and still report correct correlation in the spectral domain. A high correlation would indicate good performance of the loop and thus negate any issues with non-linearity. A low coherency would be a cause of concern and would need to be diagnosed carefully to arrive at root-cause analysis of the oscillations.

Simulated examples:

Simulation model 1: A Simulation model using Matlab/Simulink is designed to excite a first order plus time delay (FOPTD) process under closed loop control with oscillations in the SP signal. This simulated data was used to detect SP activity.

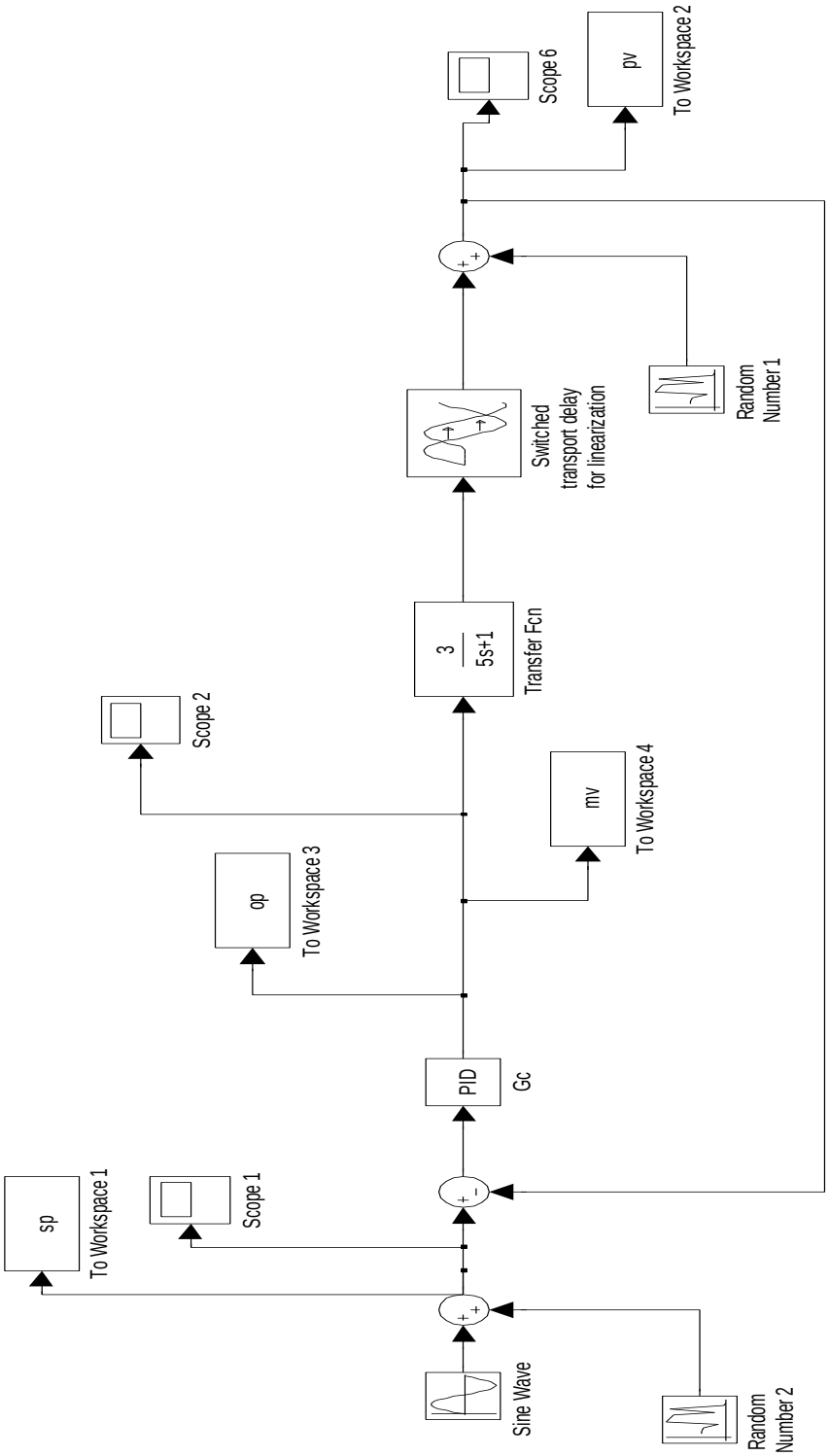


Figure 2.8: Simulink block diagram without any non linearity in the loop

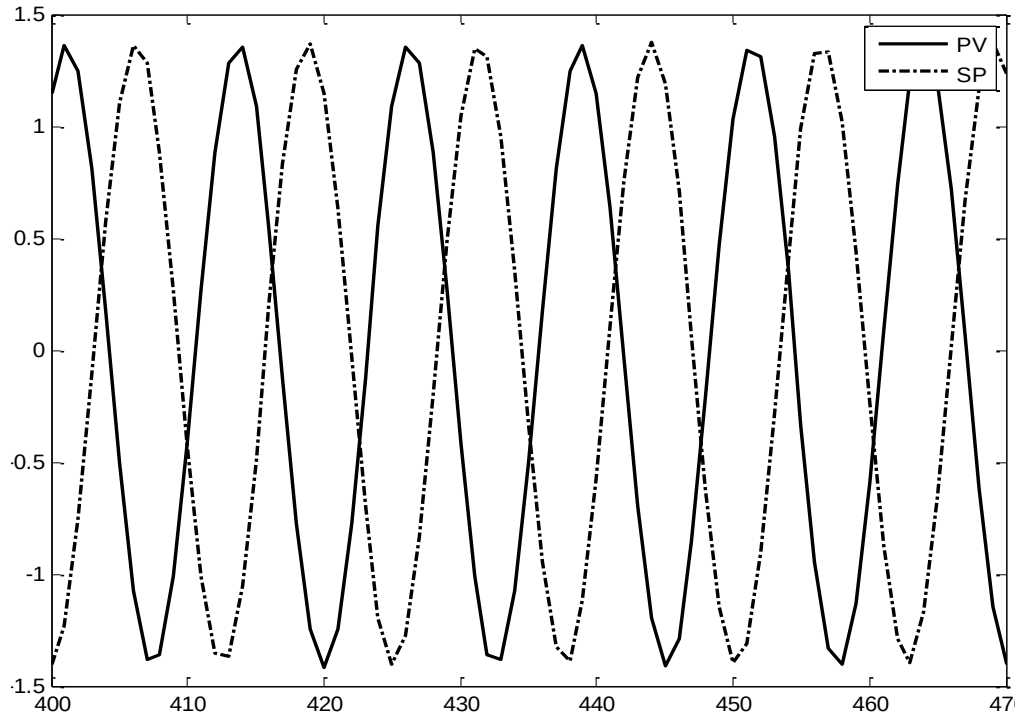


Figure 2.9: Zoomed plot of PV and SP from simulation model as shown in Figure 2.8

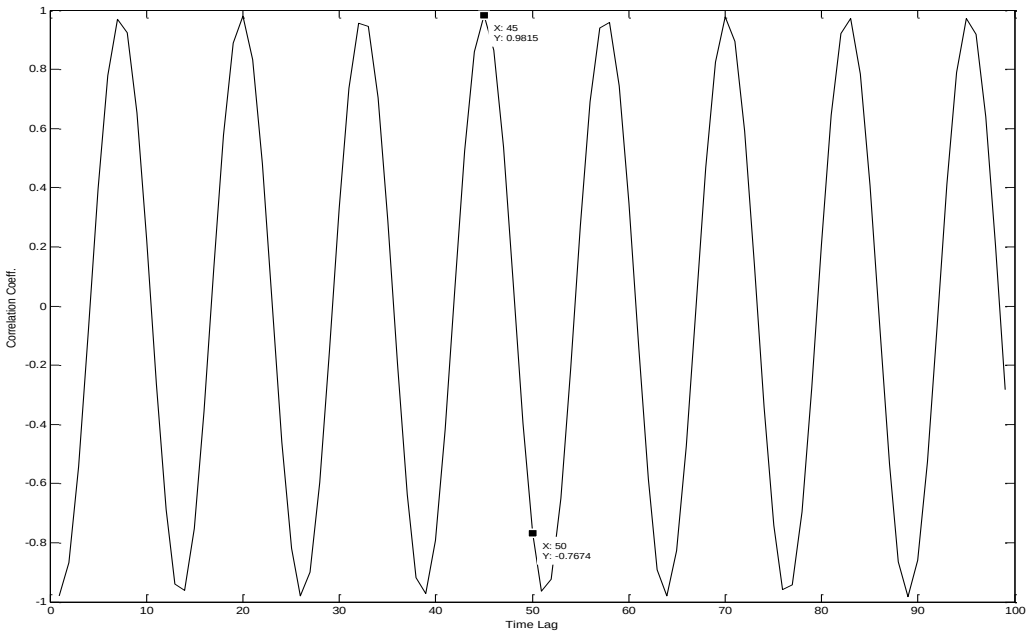


Figure 2.10: Correlation of SP and PV at different time lags from simulation model 1 in Figure 2.8

From Figure 2.10, the present algorithm fails to check this case as a 'high' SP activity case, even though the SP signal has significant number of zero crossings and the correlation between the SP and PV, without taking the time lag into account, is smaller than the threshold value of 0.8. We can observe that correlation between PV and SP is maximized at time lag = 5 (correlation= 0.9815) but in the current algorithm, we proceed with the correlation calculated at time lag=0 (to yield a correlation of 0.7674), thus resulting in a false alarm. With the suggested amendment in set point activity check, we are more confident in inferring this as an oscillation in loop because of SP oscillation.

Simulation model 2: In Figure 2.11, a simulation model is designed to excite a FOPTD process under closed loop control with non-linear oscillations from another interconnected loop. Consider, control **loop 1 with the valve stiction block** and **loop 2 without valve stiction** *i.e.* loop 1 is excited with step signal and PV of loop 1 is SP for loop 2.This simulated data from both the control loops were used to detect SP activity and to check that there would be no false alarms in the loop with no stiction.

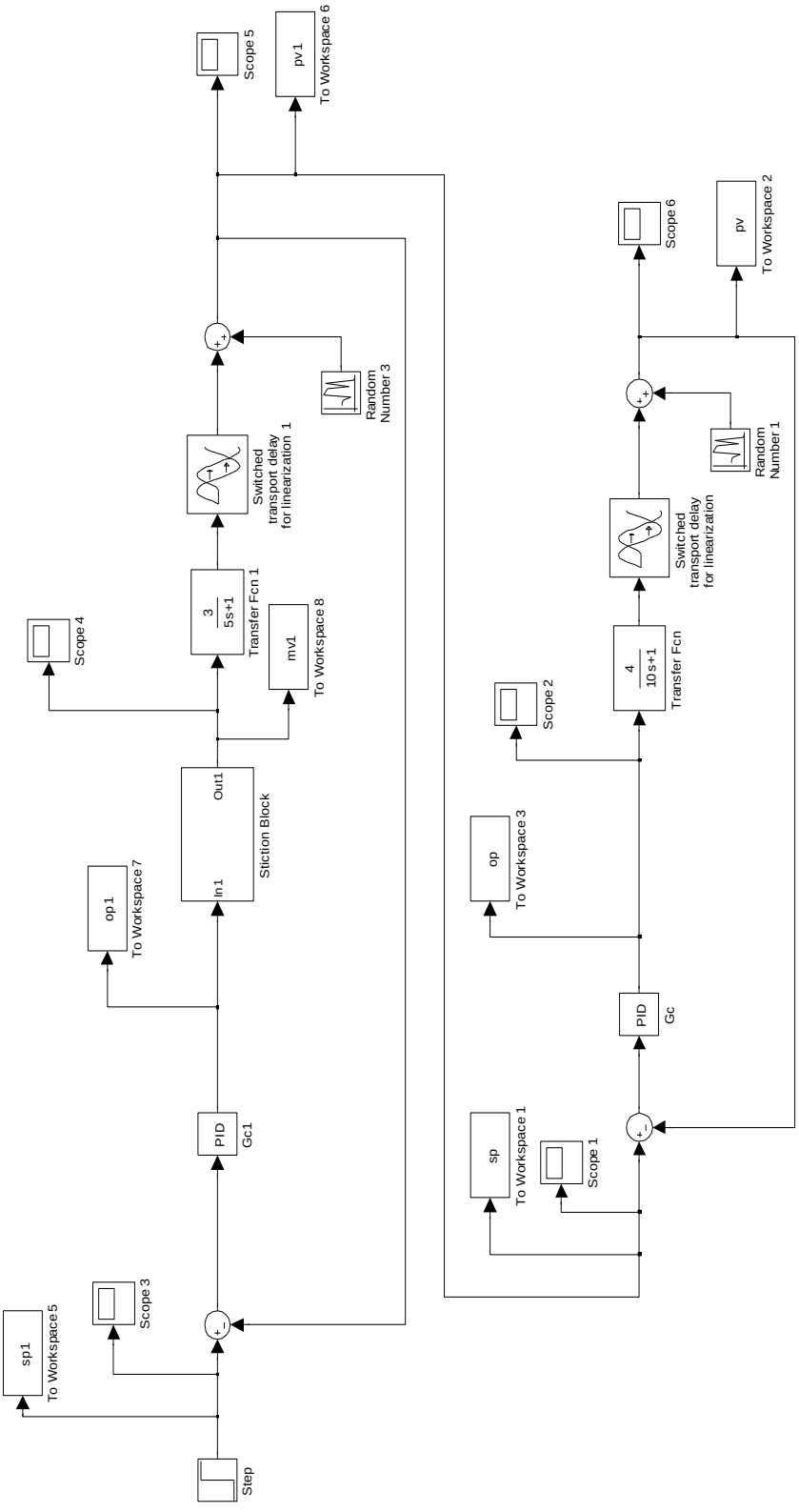


Figure 2.11: Simulink model of inter-connected loops

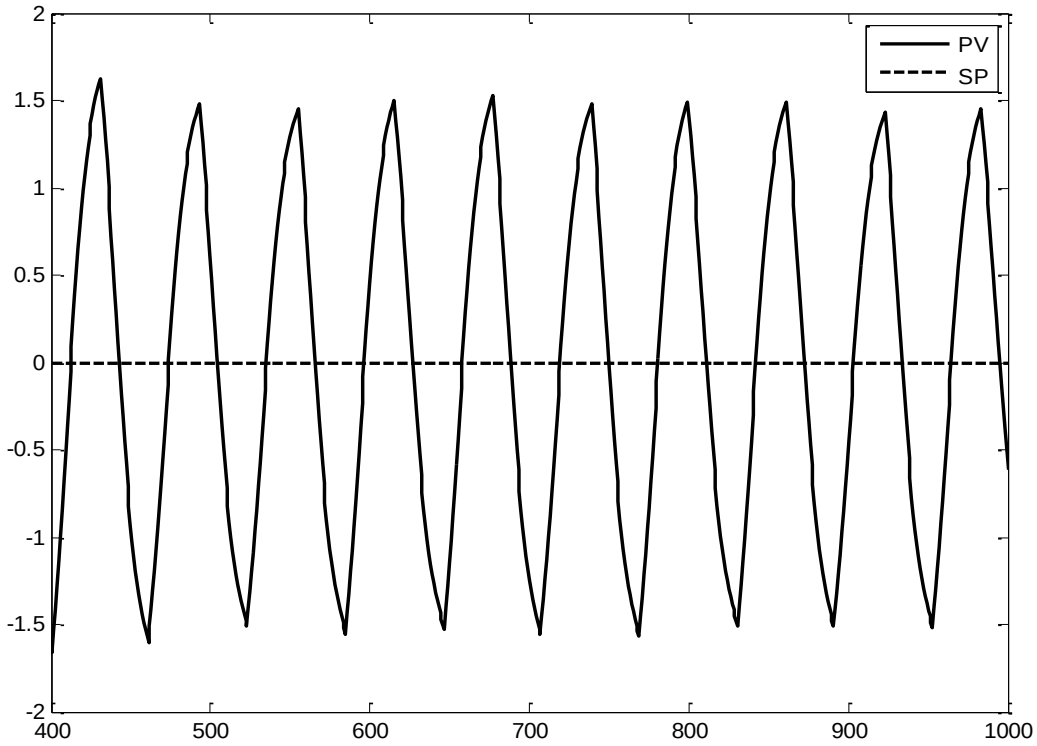


Figure 2.12: Zoomed plot of PV and SP for loop 1 (with stiction)

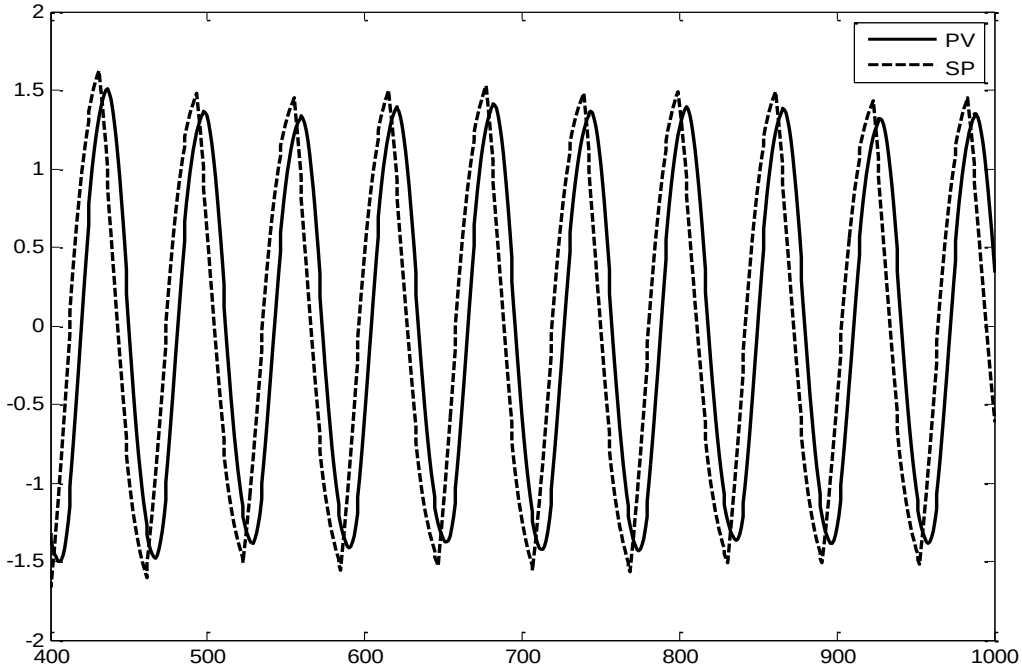


Figure 2.13: Zoomed plot of PV and SP for loop 2 (without stiction)

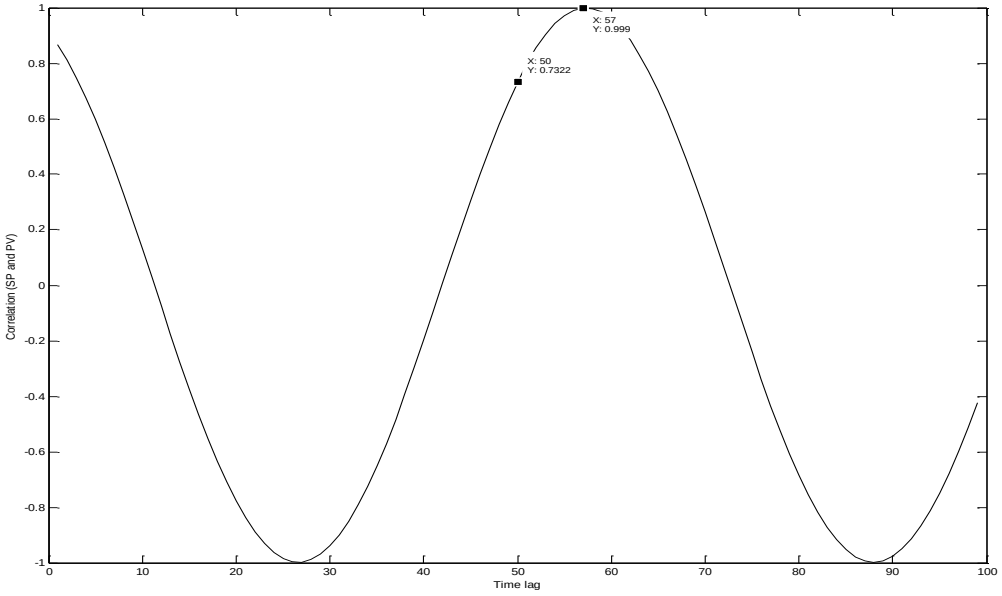


Figure 2.14: Temporal correlation of SP and PV for loop 2 at different time lags

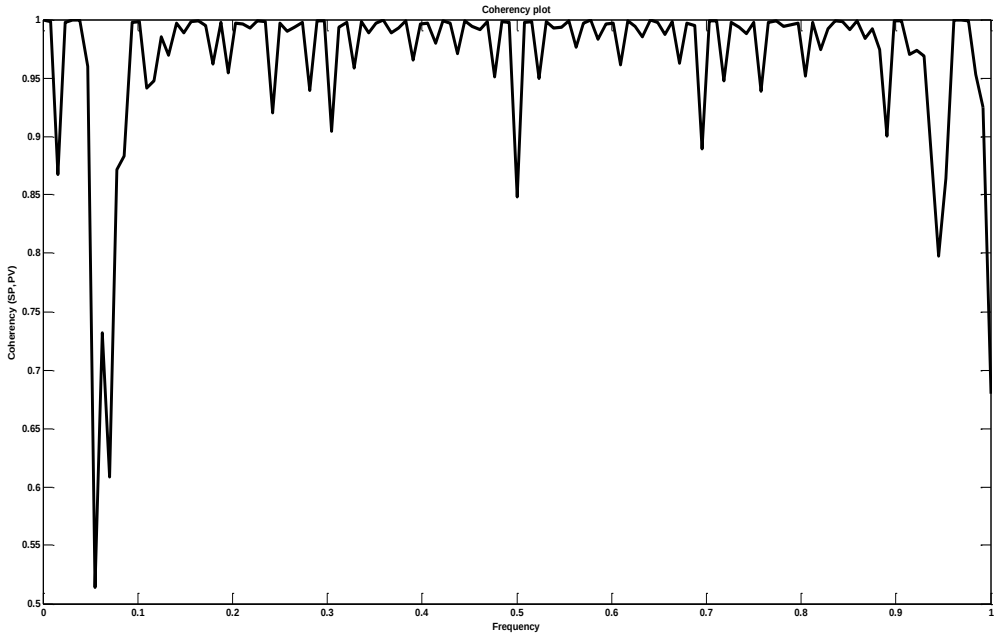


Figure 2.15: Coherency of SP and PV for loop 2

The first loop has stiction with little or no SP activity. Both algorithms are able to correctly detect stiction in this loop (see Figure 2.12).

For the interconnected second loop it can be seen from Figure 2.14, that the present algorithm fails to detect this as a 'high' SP activity case, even though the SP signal has a significant number of zero crossings and high NLI; the correlation between the SP and PV without taking the time lag into account, is smaller than the threshold value of 0.8. We can observe that the correlation between PV and SP is maximized at time lag of 7 (correlation= 0.999) but the current algorithm calculates the correlation at time lag 0 (correlation= 0.7322), thus reporting this as a case of false alarm, *i.e*. reporting stiction when there is none.

Simulation model 3: The simulation model represented in Figure 2.16 represents a common case of cascaded loop. The revision suggested in SP activity check will be able to capture stiction in the cascaded loop, as the simulation in Figure 2.16 suggest inner loop has sticky valve. As we know inner loop will have faster dynamics compared to outer loop thus control valve will exist in inner loop. If revised SP activity check detects valve problem in the loop, the method will report valve problem in inner loop.

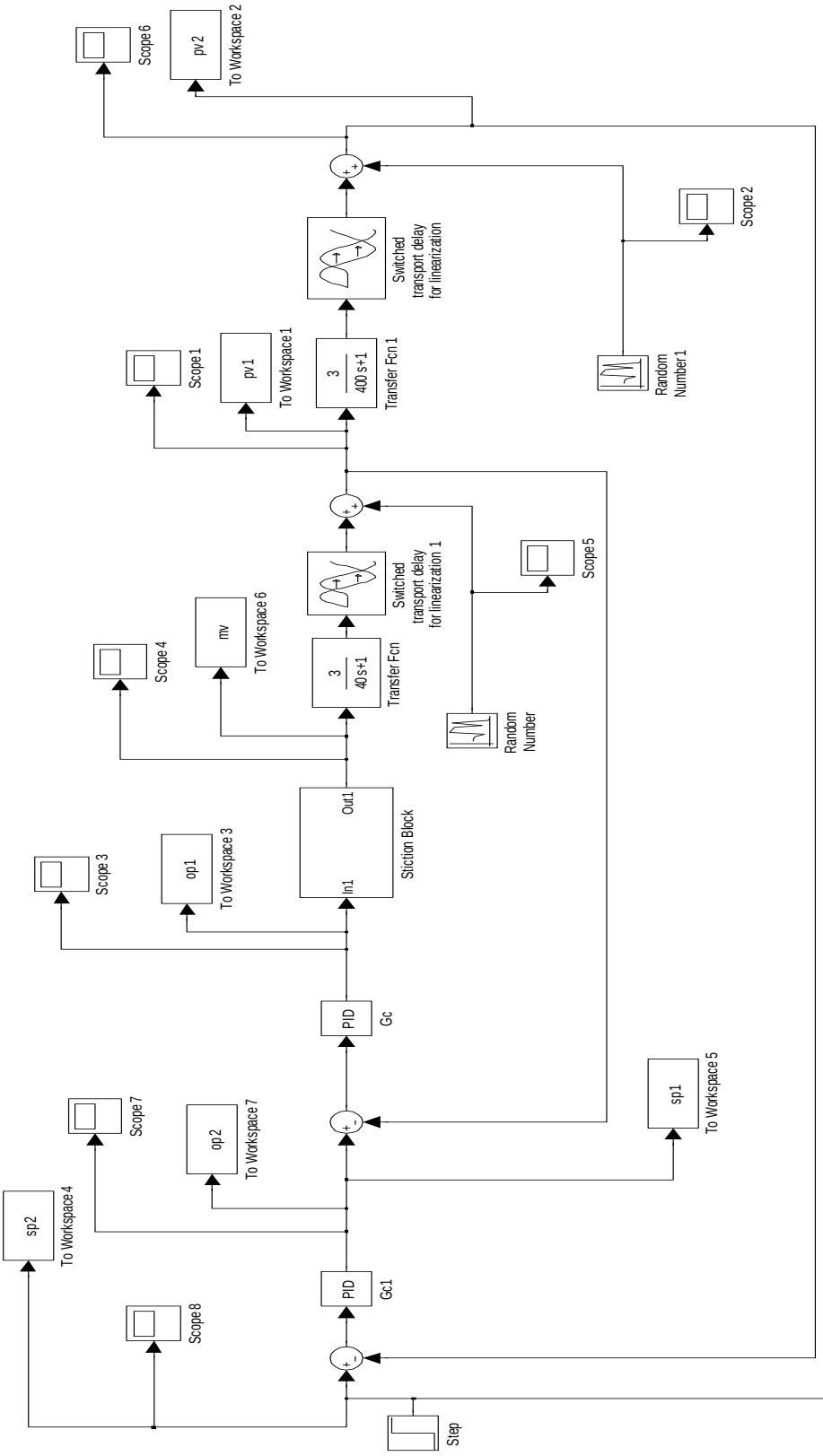


Figure 2.16: Simulink model of cascaded

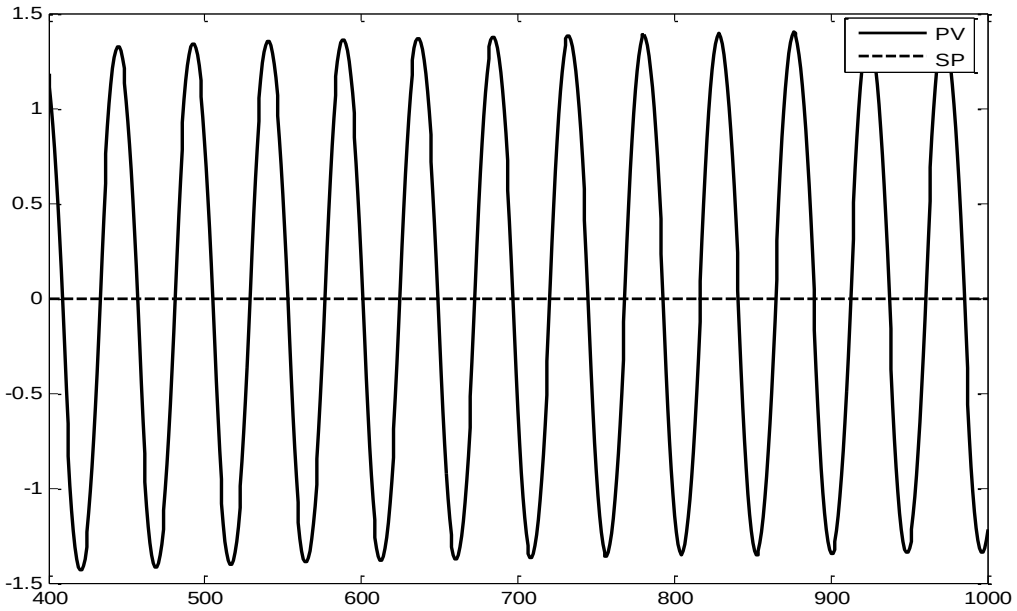


Figure 2.17: Zoomed plot of PV and SP for master loop

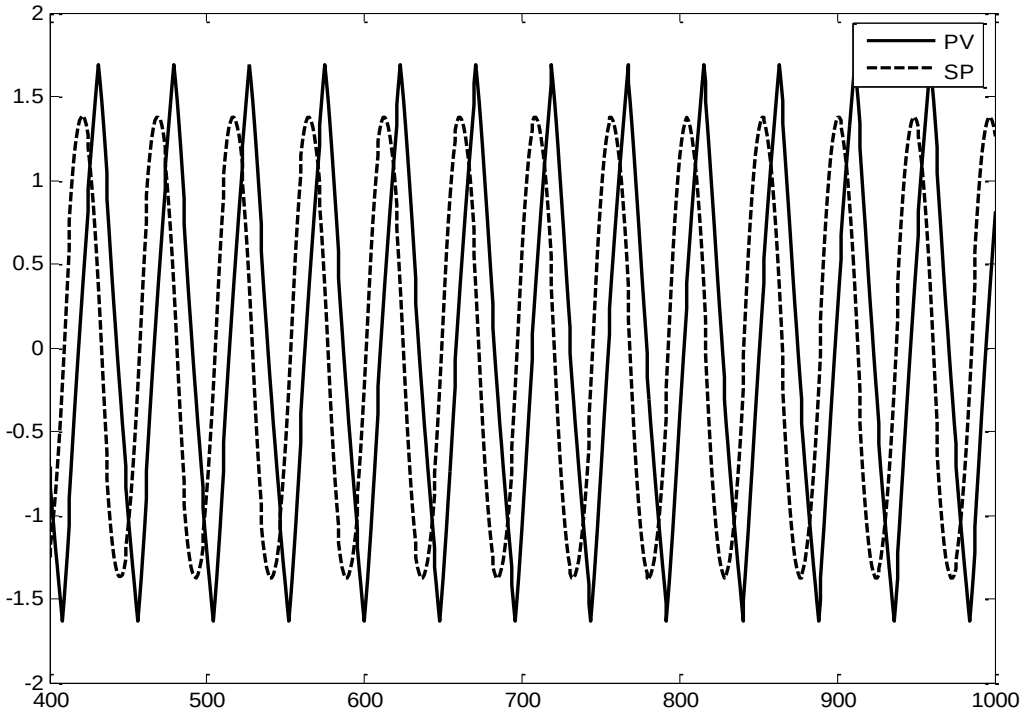


Figure 2.18: Zoomed plot of PV and SP for slave loop

To minimize the number of false alarms due to SP activity, two methods are suggested:

1) **Coherency plot:** From Figure 2.15, we can observe that because of high correlation between SP and PV, the magnitude of squared coherence is consistently high, thus an **appropriate threshold** can be designed to capture the high correlation between **PV and SP**. This method is advantageous over temporal correlation analysis because spectral correlation of coherency analysis is phase blind requiring less computational effort.

2) **Time lagged correlation:** From Figures 2.10 and 2.14, we can observe that correlation is function of time lag and is maximized at particular time lag. So as an alternative to coherency analysis between PV and SP, we suggest calculating correlation at different time lags and then searching for the maximum correlation, and subsequently using this maximum correlation for further analysis. This will require additional computational resources.

These changes were applied and checked on the industrial dataset shown below. In both cases, segments of width 2048 points with moving overlapping width of 200 points were selected:

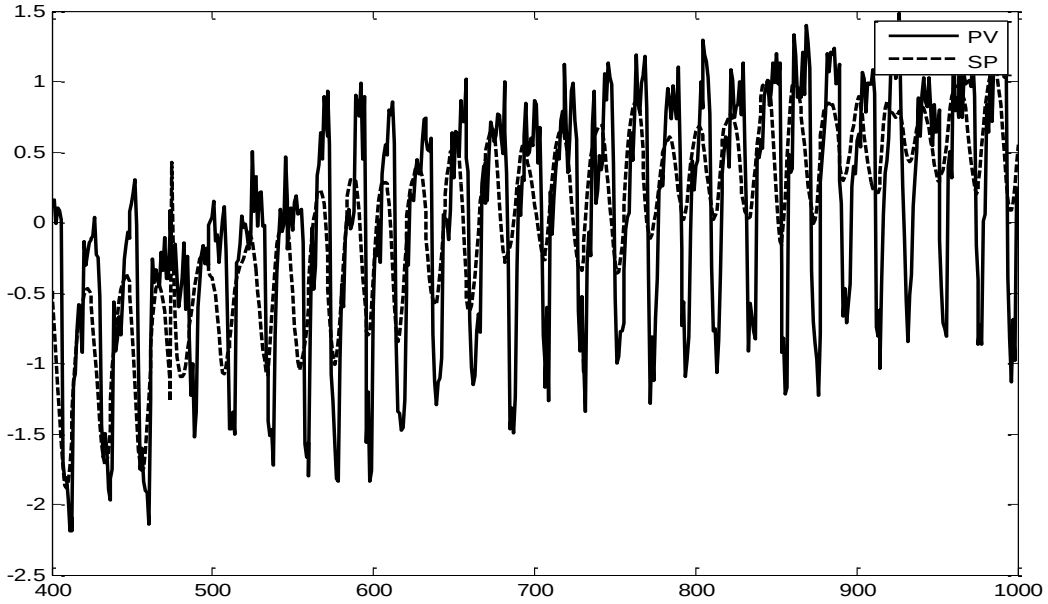


Figure 2.19 ***(Tag FC loop CHEM 14 Ref [8] page number 298)*****: This case was correctly diagnosed as 'high SP activity'**

2.4 Brief outline of algorithm:

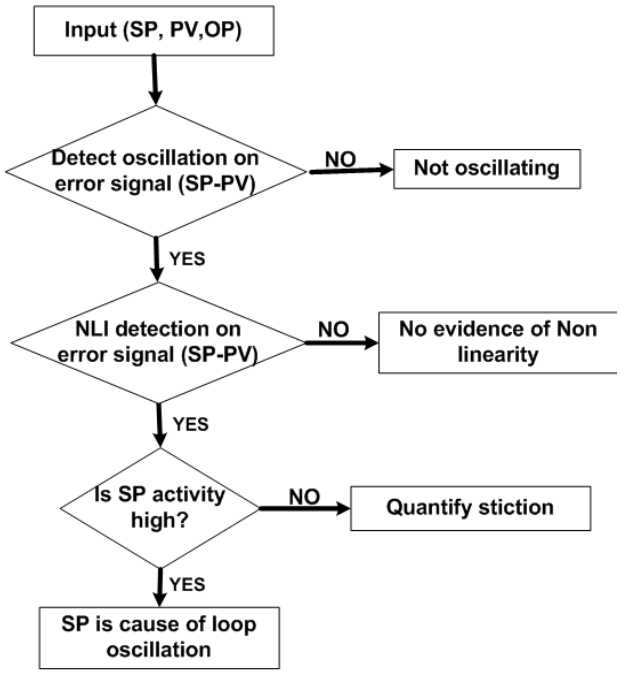


Figure 2.21: Flow diagram of the revised algorithm

The above flow chart gives a brief outline of the algorithm for valve stiction quantification. The steps followed in the algorithm are explained in detail below:

1) Input (SP, PV, OP)
2) Detect oscillations from **detrended** error (SP-PV) data on segmented overlapping windows as present in the current algorithm.
3) If oscillations are detected, NLI calculations are performed on **smaller overlapping windows** of error signal as suggested earlier in this report.
4) If NLI detected to be high for more than half of the segmented windows, a SP activity check is performed as **suggested above**. We also suggest reporting percentage of data for which high NLI is detected.
5) If SP is not the cause of loop oscillation, quantify stiction on larger data as done in current algorithm

Output: After thorough analysis of PV, SP and OP data the algorithm reports the following five parameters with changes implemented in second and fifth output parameters to make the diagnosis of the loop easier.

1) Stiction
2) **Maximum NLI**
3) Error due to stiction
4) Stiction confidence
5) **Percentage of window segments with 'high' NLI**

2.5 Concluding remarks:

The extensive data analysis reported below on some benchmark industrial data (*Ref [8] page number 298-366)* set reveals that the extensions and improvements suggested has significantly enhanced the diagnosis process. The diagnosis is not

only restricted to stiction problems but with these recommendations the algorithm can also address 'other' valve problems. The success rate for stiction detection and quantification was increased from **70%** to more than **90%**.

Notation:

NS = No Stiction

MA = Missed alarm

S = Stiction

NLI = Non Linearity Index

Color notations:

Blue: Represents cases of missed alarm with the revised method as described above.

Red: Represents cases of missed alarm with current algorithm but were diagnosed correctly with revised algorithm.

Gray: Represents cases of false alarm with current algorithm but were diagnosed correctly with revised algorithm.

Results of case study 1: All of the changes as suggested above were implemented in the MATLAB code and tested on simulated and industrial datasets. Table 2.1 shown below gives comparison of the results from current MATLAB code and revised MATLAB code applied to the industrial dataset (24 hrs data sampled at rate of 1 minute). The comparison of the results below show fewer cases of false and missed alarms (marked in red). We suggest that industrial engineers try this out on a larger data set and recommend appropriate threshold on the coherency level.

Current algorithm			Revised algorithm	
Loop name	Stiction	NLI	Stiction	NLI
FC190A	0.62	0.25	0.89	0.54
FC314	0.00	0.30	0.00	0.63
FC566	0.92	0.23	0.11	0.31
FC977	NS	NA	1.04	0.89
FC308A	NS	NA	2.70	0.75
LC010	2.86	0.55	2.67	0.87
LC093	0.02	0.92	0.06	0.97
LC156	NS	NA	0.33	0.08
LC164	NS	NA	0.17	0.65
LC168	0.70	0.30	0.72	0.59
PC103	2.80	0.98	2.76	0.99
PC112	3.69	0.24	4.77	0.87
PC134	NS	NA	0.24	0.68
PC268	3.40	0.66	4.02	0.87
PC950	NS	NA	1.95	0.52
PC018	7.64	0.47	2.02	0.94
PC008	0.32	0.83	NS	NA
PC009	0.53	0.76	0.52	0.64
PC010	0.48	0.95	0.60	0.95
PC028	0.26	0.38	NS	NA
TC061	1.40	0.37	1.04	0.60
TC063	1.22	0.23	0.94	0.56
TC106	0.31	0.69	0.21	0.94
TC007	2.18	0.68	1.72	0.84
TC015	0.32	0.77	0.31	0.84

Table 2.1: Current algorithm (5 cases of missed alarm (marked in yellow) and 2 cases of false alarm (marked in green)), compared with revised algorithm (5 cases which were missed with the earlier code (marked in red) and 1 case of false alarm (marked in blue)). To make this comparison easier the cases of false alarms (marked in orange and grey) are distinguished in both Tables.

Results of case study 2: The changes proposed in the bi-coherence based algorithms have been shown to reduce the number of missed and false alarms. We can observe that there are fewer cases of missed alarm with the revised algorithm

and most of these missed cases can be attributed to failure of oscillation detection algorithm. The stiction detection algorithm can thus be further fine-tuned by having a robust oscillation detection algorithm. The brief summary of results from the analysis is compiled in Table 2.1 and Table 2.2:

Chemicals		
	Result	**Remarks**
Loop 1	29 % High	This is a case of missed alarm with current algorithm
Loop 9	NS	MA
Loop 14	NS	This is a case of false alarm with current algorithm. Case of High SP activity.
Loop 15	NS	MA
Loop 16	40 % High	This is a case of missed alarm with current algorithm
Loop 17	NS	MA, no oscillations detected
Loop 18	NS	This is a case of false alarm with current algorithm. Case of High SP activity.
Loop 24	NS	This is a case of false alarm with current algorithm. Case of High SP activity.
Loop 27	S	This is a case of missed alarm with current algorithm because no NL detected
Loop 30	45 % High NLI	This is a case of missed alarm with current algorithm
Loop 32	NS	This is a case of false alarm with current algorithm. Case of High SP activity.
Loop 35	30 % High	This is a case of missed alarm with current algorithm
Loop 37	NS	MA
Loop 41	NS	MA
Loop 43	NS	MA, no oscillations detected
Loop 49	NS	MA, no oscillations detected
Loop 51	20 % High	This is a case of missed alarm with current algorithm
Loop 70	40 % High	This is a case of missed alarm with current algorithm
Loop 72	20 % High	This is a case of missed alarm with current algorithm
Loop 73	NS	MA
Loop 75	20 % High	This is a case of missed alarm with current algorithm

Table 2.2: Results from chemical plat data (*Ref [8] page number 298)*

Table 2.3 reports the combined analysis of the two case studies reported above and shows a significant improvement in the success rate for stiction detection.

Total cases = 101	Current algorithm	Revised algorithm	Percentage improvement
Missed alarms	22 %	8 %	14 %
False alarms	6 %	1 %	5 %
Successful detection	72 %	91 %	19 %

Table 2.3: Summary of results from the above two case studies for a total of 101 loops. The overall success rate for the revised algorithm improved from 72% to 91%.

Chapter 3

Complete Solution: Automated, Unsupervised Detection of Malarial Parasites in Microscopic Images

Introduction 3.1

The war between homo sapiens and the mosquito borne parasite plasmodia has been epic. The results are so bitter that we are loosing a million lives [19] out of more than 250 million infections annually. The overall impact of the disease around the world is massive and it is an enormous burden to developing countries. Figure 3.1 gives an idea how some parts of the world have been affected by the disease. In recent reports the tentative death tolls are higher than those reported by World Health Organisation (WHO).

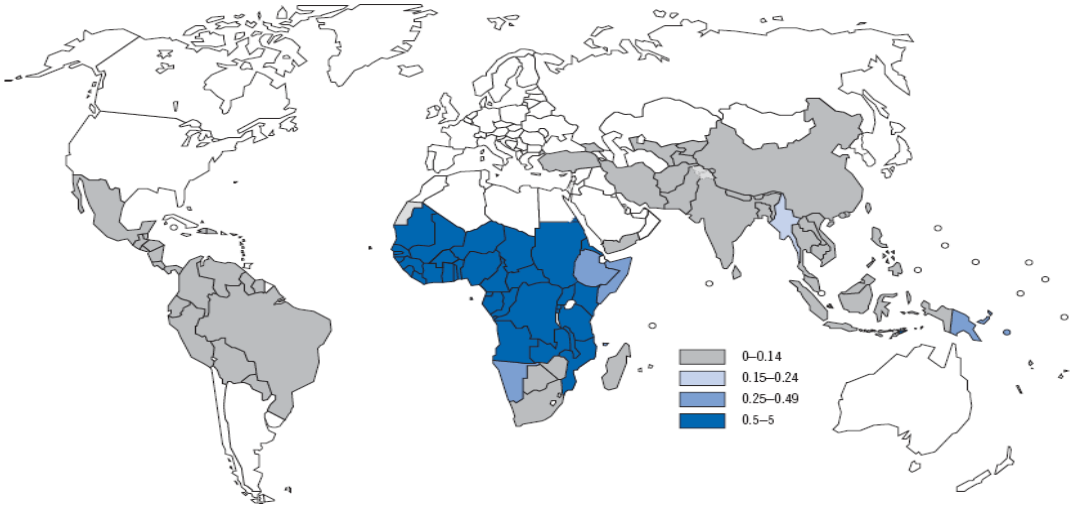


Figure 3.1: Estimated deaths from malaria per 1000 population, 2006

There is an ambitious goal to eradicate malaria with-in a decade and to improve overall global health. Money funnelled into malaria control has soared to over $ 1 billion a year [6]. The malaria control program can roughly be categorized under three major steps, preventing the disease to spread by providing insecticidal nets, spraying insecticides, speedy and low cost diagnosis and cheap and effective medication. Doctors in the developing world often lack laboratory facilities to authenticate cases of suspected malaria. Perhaps more often, they never even get to see patients who have the disease — many patients either cannot afford the time or money to see a doctor or they simply self-diagnose and take cheap over-the-counter medications to battle malaria-like symptoms [6].

To better understand why malaria has become such a threat and what can be done to stop the disease, it helps to know a little biology. Malaria is caused by four closely related parasites *Plasmodium vivax, Plasmodium ovale, Plasmodium falciparum* and *Plasmodium malariae*, the deadliest of which is *Plasmodium falciparum*, which has a particular fondness for anopheles mosquitoes. The parasites enter the bloodstream when an infected mosquito bites a human. Then they multiply inside the host's liver and red blood cells. Eventually the red blood cells (RBCs) burst with a new generation of parasites, causing fever, shivering, pain and sometimes death. The cycle of transmission is complete when another mosquito bites an infected person and picks up more parasites.

There are various techniques [5] to diagnose malaria of which manual microscopy is considered to be the gold standard. Manual microscopy has advantage over other techniques that it is sensitive and specific. However, one of the disadvantages of diagnosis using manual microscopy methods is that it requires extensive human intervention during the diagnosis process which can often lead to late and sometimes erroneous diagnosis. The operator requires extensive training to gain expertise in the diagnosis using a microscope and in any case because of the sheer volume of the samples that need to be analyzed, the method is not consistent. The focus of this research is to upgrade the diagnosis process such that

it could be fast, reliable and low cost with minimum human reliance. Such a diagnosis system can be designed by understanding the procedure and standards followed by experts.

3.2 Manual microscopy

The WHO practical microscopy guide for malaria provides detailed procedures for laboratory practitioners [20]. Diagnosis initially requires observing blood smeared slide under the microscope for detecting the presence or absence of parasites. The practitioners generally observe two types of slides using fast Giemsa stain protocol to highlight the parasites. The two types of blood slides are thick smeared and thin smeared slides as shown in Figure 3.2, where the violet coloured dots with in the RBC can be identified as stained parasites. The disadvantage of Giemsa staining is that it also stains other artefacts such as white blood cells, platelets, dirt etc. This problem of other stained objects needs to be considered carefully when comparing results of automated image-based diagnosis with manual microscopy.

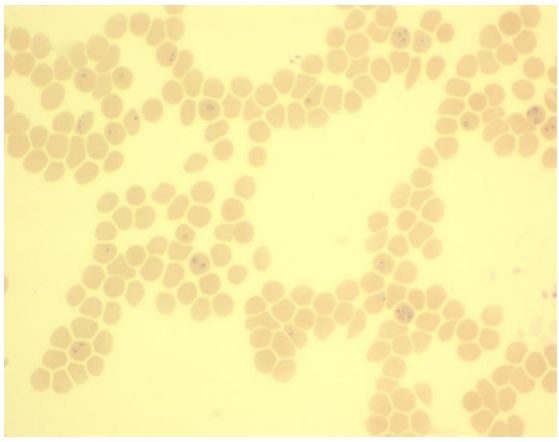


Image of a thin blood smeared slide

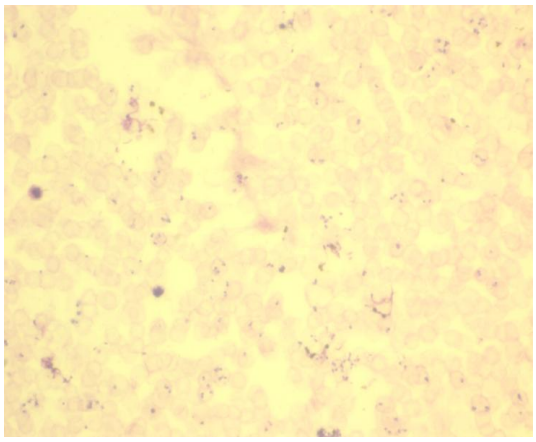


Image of a thick blood smeared slide

Figure 3.2: Images of thin and thick Blood smeared slides

The probability of detecting parasite in thick blood smeared slides is higher because of the larger volume of blood observed. If the expert suspects the presence of Malarial parasite in thick blood smeared slides then the diagnosis process is followed by examination of a thin smeared blood slide. The thin blood smear slide is used for the enumeration of the infection i.e. determining the number of infected RBCs per 1000 RBCs. The process of enumeration requires manual count of the number of RBCs in the observed microscopic field followed by counting of number of infected RBCs. Apart, from enumeration pathologists also need to report the life cycle and the specie causing the infection; thus the process of diagnosis is extensively dependent on experts.

3.3 Materials

Anonymized thin blood film images acquired as described in Figure 3.3 were obtained from Alberta Health Services laboratories at the University of Alberta. The samples obtained mostly had low number of parasite(s) in early stages (ring trophozoites) of their life cycle. Such features are often hard to detect. In addition several samples did not have any parasites (negative controls). The samples were stained using a fast Giemsa protocol to highlight the parasite(s) and were initially examined by hematopathologists with expertise in malaria diagnosis. Slide images were acquired using a charge coupled device (CCD) camera with different range of magnification. Some images had variable stain characteristics making computer based detection more challenging. In total 10 patients samples were examined, parasitemia was reported as number of infected RBCs per thousand RBCs. Thus for each patient, 10 images each representing one microscopic field at 1000X magnification and containing approximately 100 RBCs were acquired to fulfill WHO recommendations. Ten negative control cases were included in a blinded fashion to check the specificity of the image based method.

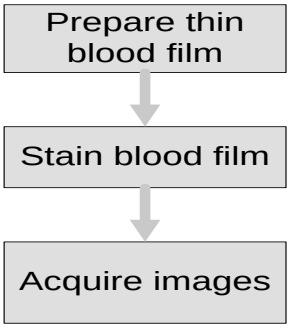


Figure 3.3: Flowchart describing the steps of image aquasition

3.4 Method

The problems of manual microscopy can be overcome by exploring computer based, specifically image-based, diagnostic methods. An automated diagnostic method can be developed by understanding the diagnostic expertise and representing it by a specifically tailored image processing based algorithm. The literature contains descriptions and details of several computer vision or image-based algorithms [12, 13, 16, 17 and 18]. However, most of these algorithms are supervised and complex, that is they need manual intervention or calibration. Considering the high fatality rate and huge volumes of samples that need to be analyzed we need a sensitive, practical and robust method with minimum human intervention. In this context, computer based diagnosis can help in the rapid, accurate and consistent identification of true Malaria cases and thus significantly reduce the financial burden on overall malaria prevention program [20] and therefore allow more resources to be assigned towards the prevention and treatment of malaria.

An automated diagnosis method can be developed by understanding the diagnostic expertise and representing it by a specifically tailored image processing based algorithm. The image processing based algorithm should perform diagnosis more or less imitating the process undertaken in manual microscopy. The algorithm should be capable of operating in an unsupervised environment and need to be robust with no false negatives. The unsupervised nature of the

proposed procedure should reduce human intervention, and in so doing speed up the diagnosis process. The algorithm should also be sensitive enough to capture parasites at all stages particularly at the early stages of their life cycle and to do this without missing any parasites irrespective of image variations. In order to perform diagnosis, the method must be capable of differentiating between parasite and artefacts. The majority of the image based diagnosis methods [12, 13, 16, 17 and 18] reported in literature do not address this requirement.

The challenge to achieve this high degree of sensitivity and robustness under an unsupervised environment has been carried out by developing a very novel and simple statistical method for image classification. The image classification problem performs the following steps: i) RBC enumeration, ii) potential parasite identification and iii) report parasitemia by counting the number of infected RBC per every 1000 RBC. The steps that constitute the image processing and image segmentation tasks are summarized in the flowchart as shown in Figure 3.4. The method was developed on the MATLAB^{TM} R2007b platform.

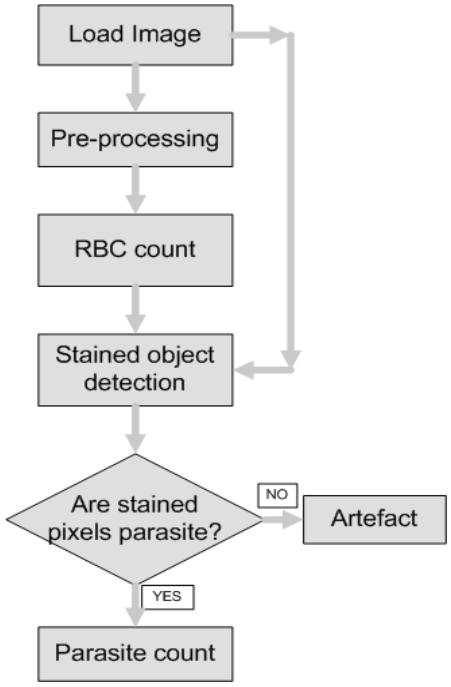


Figure 3.4: Flowchart describing the steps of the proposed algorithm

3.4.1 Pre-processing:

The purpose of pre-processing is to remove unwanted objects and noise from the image to facilitate image segmentation into meaningful regions. The image processing enhance the input image for visual purposes. The series of steps required to carry out image pre-processing are as follows:

i) Load colored (Red, Green and Blue channel) or gray scale image, the colored image is first converted to a gray scale image. The contrast of the gray scale image is enhanced using local histogram equalisation [11-12] to enhance the visibility of the parasites and RBCs.

ii) The next and important step in image segmentation is to extract meaningful regions or in other words distinguish objects from the background. The common way described in the literature is to use edge detection algorithms [3 and 4]. Edge detection is a method of identifying pixels in image, where the intensity value changes sharply. An example of edge detection is shown in Figure 3.5.

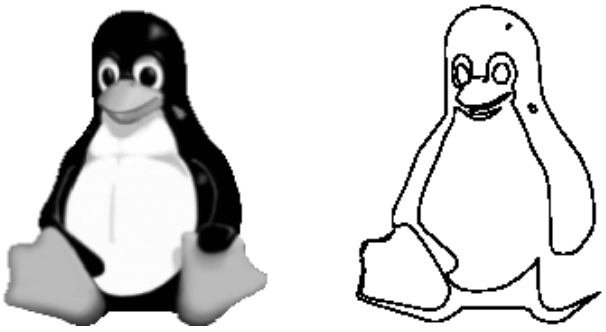


Figure 3.5: Example of edge detection

Edge detection algorithms use gradient information followed by smoothing and morphologoical boundary closing. It is not always easy to detect edges and the location of an object in presence of substantial noise. This makes edge detection algorithms computationally exhaustive and less sensitive. The novel and robust approach to detect a closed curve in an

image is to detect regions with statistical similarity inside and outside the curve. We require a method which can overcome the problems of edge detection algorithm Chan & Vese et al. [2 and 9] proposed energy minimization of the image to detect objects embedded within an image. The method is independent of gradient information and relies on the region statistics inside and outside the curve.

The goal of implementing the Chan-Vese based boundary detection algorithm to segment an image into meaningful regions, in our case separate RBC and artefacts from the background is shown in Figure 3.6. The binary label for statistically similar regions is represented in Figure 3.7.

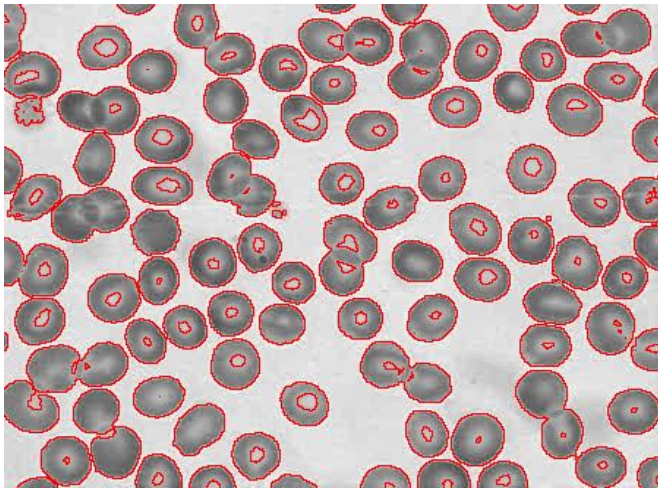


Figure 3.6: Boundary extracted image using the Chan-Vese segmentation method

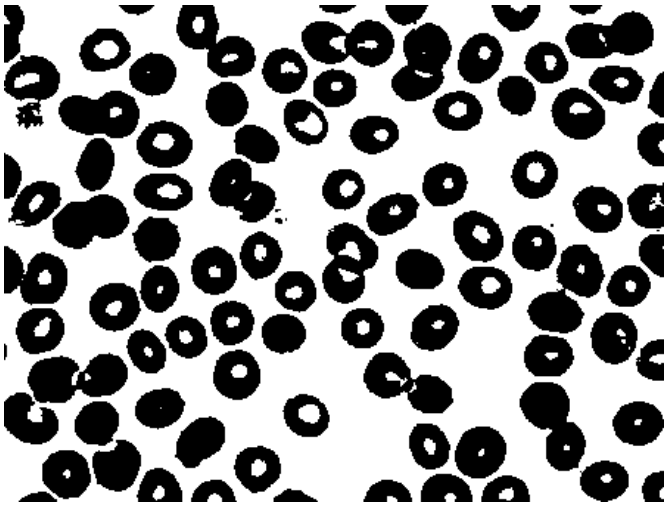


Figure 3.7: Binary mask from the Chan-Vese segmentation method

iii) The binary image of statistically similar region generated after the Chan-Vese segmentation is able to distinguish RBCs from the background. However because of the biconcave shape of the RBC as shown in Figure 3.8, the central pallor is assigned the same features as the background, as shown in Figure 3.7. Because of this we need to remove the holes in the binary image to have a good RBC count. To perform this task a 'hole filling' algorithm was designed as described with a simple illustrative example in Figure 3.10.

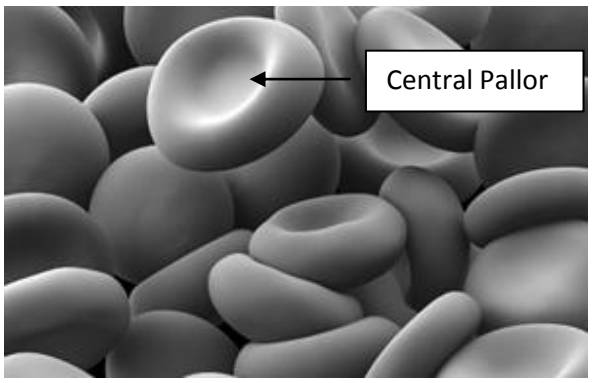


Figure 3.8: Bi-concave shape of RBC

The technique to fill the holes in the binary digital image is to distinguish the background from the central pallor. Connected component analysis can be used to separate the background from the central pallor in the binary image. Connected component analysis extracts the information on pixel connectivity in 2-dimensional image by labelling connected pixels possessing the same intensities. Figure 3.9 shows an example of 8 sided connectivity. All the connected components were extracted shown as in Figure 3.10b and the mathematical operation of intersection was performed between the largest connected component with the original binary mage. The idea used was that generally in an image background is the largest connected component. The final hole filled image is shown in Figure 3.10c after the operation of intersection. The hole-filled image is subjected to minute erosion [4] using disk-shaped structuring element of radius 5 pixels, the resulting image after hole-filling and erosion is shown in Figure 3.11. The erosion removes any isolated pixel(s) in the image thus reducing the number of artefacts.

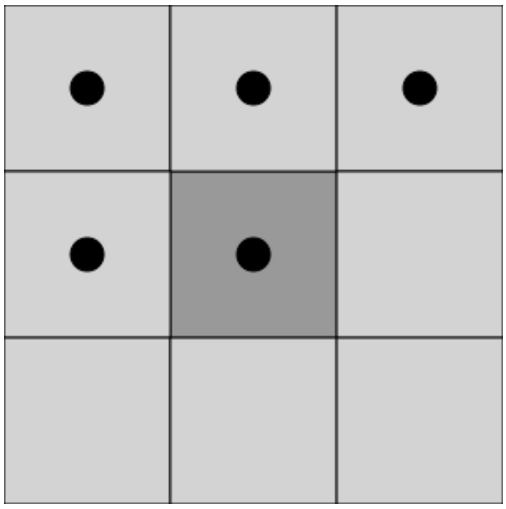


Figure 3.9: 8-way connected component

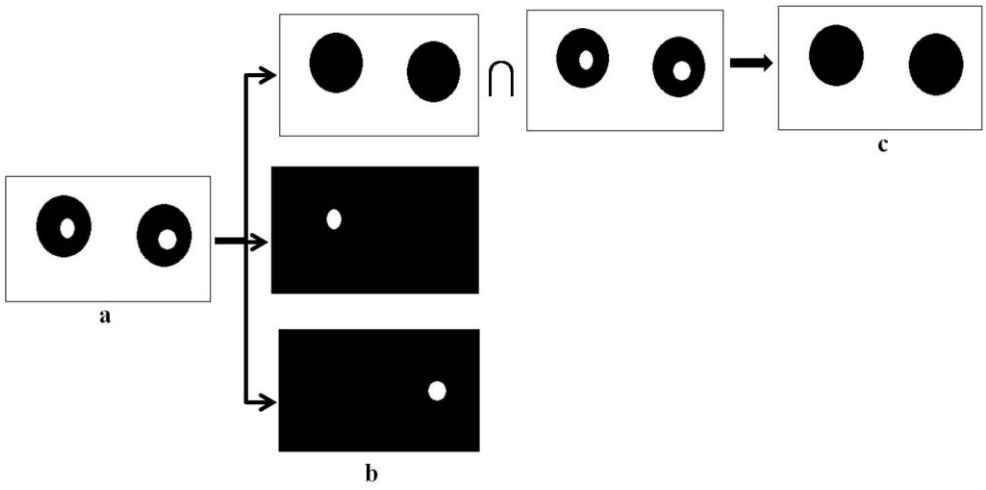


Figure 3.10: Example to illustrate the 'hole-filling' algorithm: a) Dummy image with holes b) All the connected components c) Holes filled image after intersection

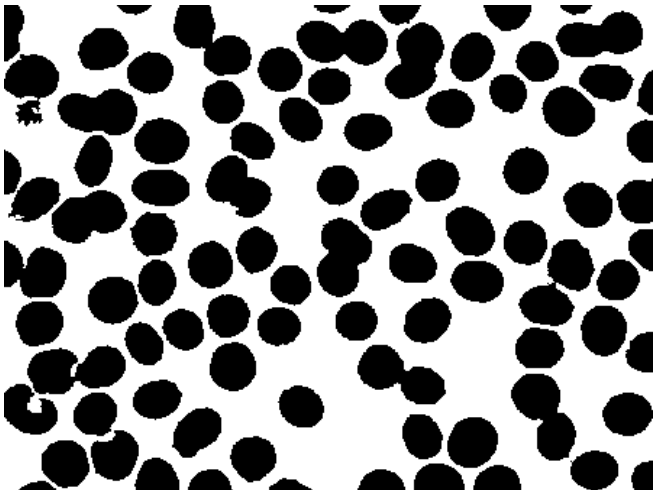


Figure 3.11: Binary image after removing spurious boundaries

3.4.2 RBC count:

The number of RBCs in the pre-processed binary image can be calculated if the total area occupied by RBCs is divided by the mean area of a RBC and then adjust the RBC count to the nearest integer. The challenge was to determine the precise mean area of one RBC for each image without any prior information (because the

size of RBCs depend on factors such as a patient's age, other blood related diseases etc). To automate this procedure in an unsupervised algorithm, we first attempt to obtain a size distribution of RBCs in pixel units. To accomplish this we observe that RBCs possess almost circular shape, this feature of RBC can be exploited and various radii circles can be fitted to the RBC to generate a circle size distribution. The mean area generated from such a circle size distribution was used to determine the total number of RBC in the image. The problem of fitting radii of various sizes to RBCs was implemented using the popular 'Hough transform' [3, 4 and 7].

Hough transform works as a powerful tool for image segmentation to extract pre-defined (line, circular, elliptical etc.) shapes in an image. The Hough transform tries to determine if the group of pixels lie on a curve of specific shape and is unaffected by image noise. The main use of the very commonly discussed Hough transform is for extracting straight lines in image. The method developed by Hough to extract straight lines can be extended to determine a circle within an image. The circular Hough transform method is a modified form of the conventional Hough transform, where the parameter equation for straight line is changed to a general equation of circle expressed in equation (1).

$$\begin{aligned} a &= x - r\cos\phi \\ b &= y - r\sin\phi \end{aligned} \tag{1}$$

Now given a gradient angle (ø) at an edge point (x, y), we can compute parameters 'a' and 'b' for pre-defined radius (r) or for given radius range. The edge points lying on the same circle are assigned to one accumulator array and this process is repeated for each point on the image. The algorithm terminates until we have assigned all edge points to one of the accumulator array thus each accumulator array possess the centre and a corresponding radius of the circle.

The modified Hough transform was applied to binary images thus maximizing the gradient of the RBCs to avoid unnecessary fitting. As can be observed from a typical slide shown in Figure 3.11, the RBCs are not all perfectly circular, which

leads to the suggestion of using the Hough transform for ellipse fitting [7]. However literature studies [3 and 4] indicate that as the number of parameters increases the algorithm becomes slow and furthermore it may have severe memory allocation problems. Therefore the Hough transform for circle of different radii was implemented and the results demonstrate the efficacy of the method in that it allowed the feature generation to be very robust and completely unsupervised. An example of this implementation is shown in Figure 3.12 followed by the corresponding circle size distribution. Table 1 represents results obtained from the analysis of one randomly picked images from patient M7.

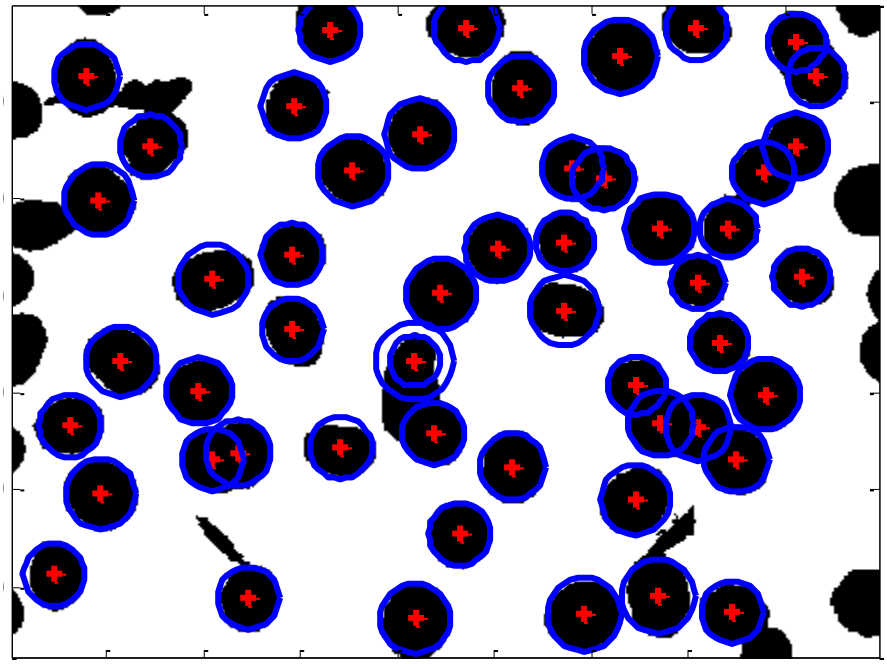


Figure 3.12: Circles of different radii are detected on binary image after removing spurious boundaries

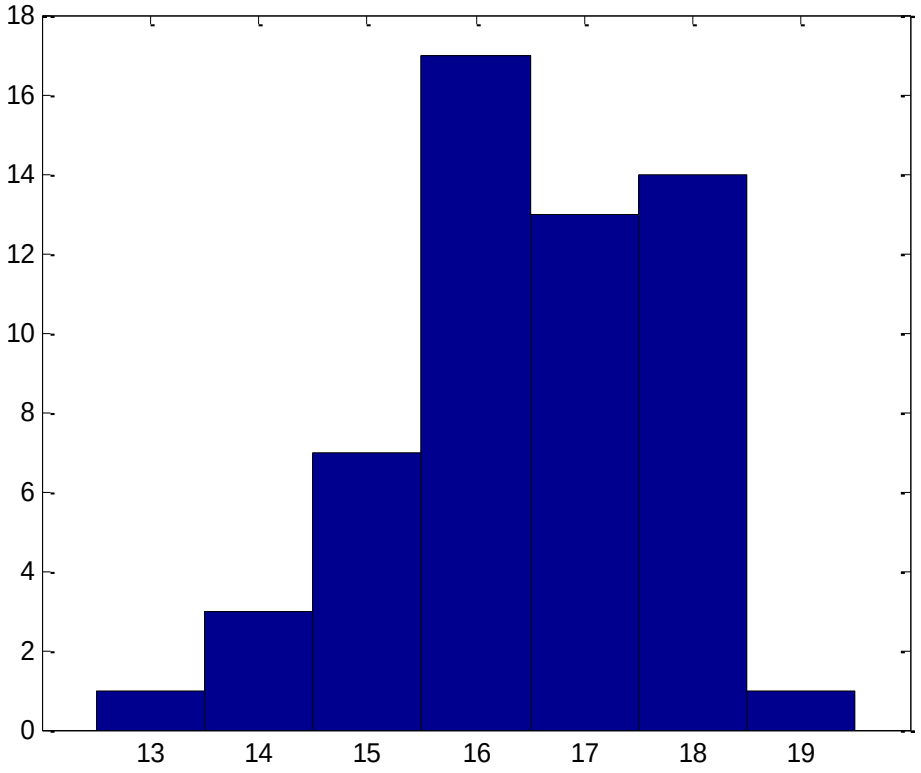


Figure 3.13: Circle size distribution

The results shown above demonstrate the efficiency of Hough transform in image segmentation. The circle size distribution shown in Figure 3.13 can be directly related to the RBC size distribution which gives a good insight of the average RBC size of the patient. The mean RBC size is an important parameter for medical experts because variation in RBC size can reflect on other blood related diseases such as anaemia. Also the method proposed above can be used as an efficient tool as an offline RBC counter. The unsupervised approach to generate RBC size distribution and RBC count acts as an added information apart from diagnosis of Malaria.

M7	Algorithm count	Manual count	Error
M7-1	67	66	1
M7-2	74	74	0
M7-3	64	64	0
M7-4	75	73	2*
M7-5	64	64	0
M7-6	76	76	0
M7-7	75	75	0
M7-8	72	71	1
M7-9	64	66	2*
M7-10	74	75	1

Table 3.1: Comparison of RBC count using the automated algorithm and its comparison with the manual count (edge cells were considered)

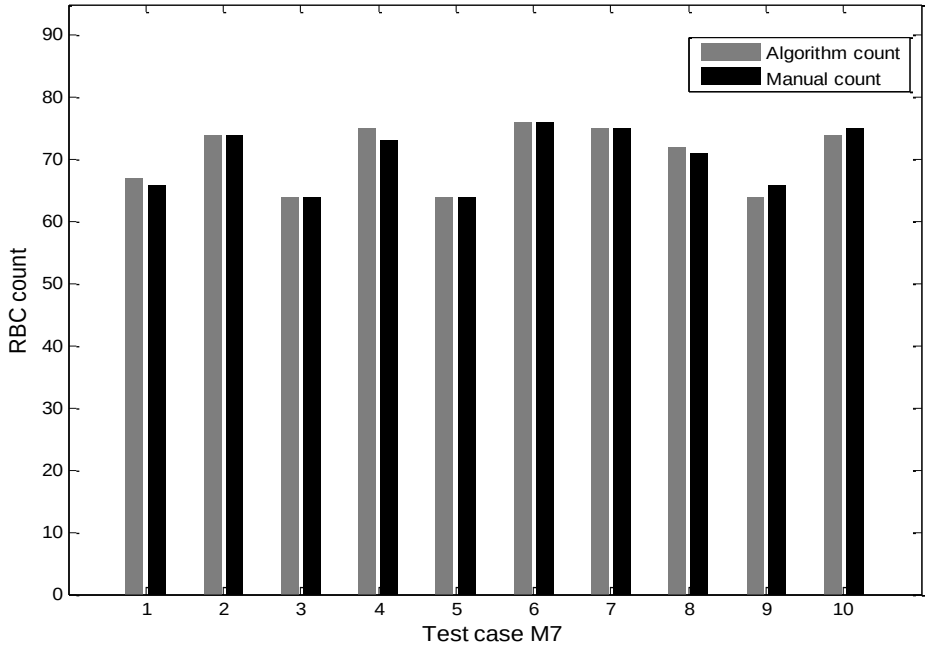


Figure 3.14: Comparison of RBC count

* Errors were due to edge cells

Manual count represented in Table 3.1 was performed by an expert practitioner. Manual count is performed as follows: an expert counts the complete RBCs embedded in the image and only counts the RBCs on two adjacent edges in order to have symmetry. Thus the error in algorithm count are generally not significant.

3.4.3 Parasite detection

The next part of the proposed methodology is the detection of potential parasites in Giemsa stained thin smear blood slide images. Considering that parasites are small stained objects embedded within the RBC, they are often distinguishable from the RBC plus other background 'image noise' and artefacts (artefacts represents platelets, WBC, dirt, dye crystals etc.) in the image. The literature is archived with several methods for image segmentation based over edge detection [12] or statistical methods using classifiers [13 and 14] for the detection of parasites. However, the edge detection method as discussed earlier is dependent on gradient information of the image and tends to generate undesired results if the parasites are barely visible or in early stages of their life cycle. On the other hand, classification method tends to overcome these problems but development of a classifier requires access to historical information making it a supervised learning method.

Our primary objective is to have a robust unsupervised method for the detection and enumeration of Malaria parasites. In this context, the property of parasites being stained can be exploited to separate them from the RBC and the background. To be able to do this we need an unsupervised, sensitive and reliable segmentation method which can extract parasite information from the image.

The visual inspection of pixel intensity data plot as shown in Figure 3.15 reveals the distinguishable clusters of background, RBC and stained pixels. The pixel intensity data plot suggests the use of a non-hierarchical method of clustering technique such as k-means clustering (KMC) [10] to segment the pixels into corresponding distinguishable groups. The advantage of clustering methods is to uniquely classify data, in this case a digital image based on an unsupervised learning methodology. These methods are also robust and can be easily implemented to achieve a desired level of image segmentation. The implementation of robust clustering method requires prior information of a number of clusters and good initial guess of cluster centroids. Commercially available software packages (e.g. MATLAB) provide KMC as a tool for data

clustering but these algorithms are not designed for image segmentation and in any case they require a good initial guess for the number of clusters and cluster centroids to have meaningful clustering. Figure 3.15 shows that the cluster of stained pixels, presumably representing parasites, has very few data points compared to RBCs and the background cluster. Clearly these giant data clusters easily overwhelm the small clusters, with the result that a clear demarcation between the RBCs and the parasites is not possible using conventional KMC.

To overcome the problems of unknown number of clusters, good initial guesses and retaining the information about small clusters in the presence of giant clusters, a modified KMC based algorithm has been developed. The problem of unknown number of clusters can be addressed as catching almost empty clusters, if we initialize the KMC with more number of clusters than expected. If the algorithm encounters an empty cluster it reduces the number of cluster by one in an iterative process until the process reaches an optimum number of separable clusters. The problem of good initial guess was combined with the initialization of the clustering algorithm in order to facilitate the clustering process by segmenting the background cluster. The background cluster can be separated from the RBC and artefacts by superimposing the hole-filled image over the original RGB image as shown in Figure 3.16. The visual inspection of pixel intensity plot as shown in Figure 3.17 reveals the background intensity suppressed to zero makes the clusters more distinguishable. The pixel intensity of the background set to zero gives us the freedom to choose the initial centroids of equal weight between zero and maximum pixel intensity value. The sparseness of the pixel map combined with a good initial guess makes the KMC algorithm fast and robust. This initial guess serves as a default guess for all cases thus rendering this as an unsupervised parasite detection and enumeration algorithm.

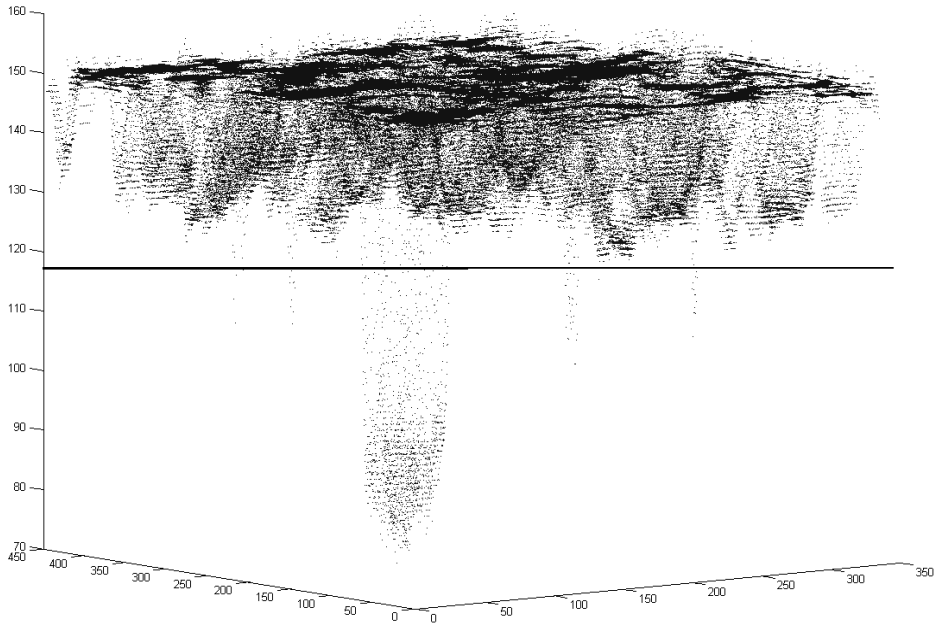


Figure 3.15: Pixel intensity plot with random threshold

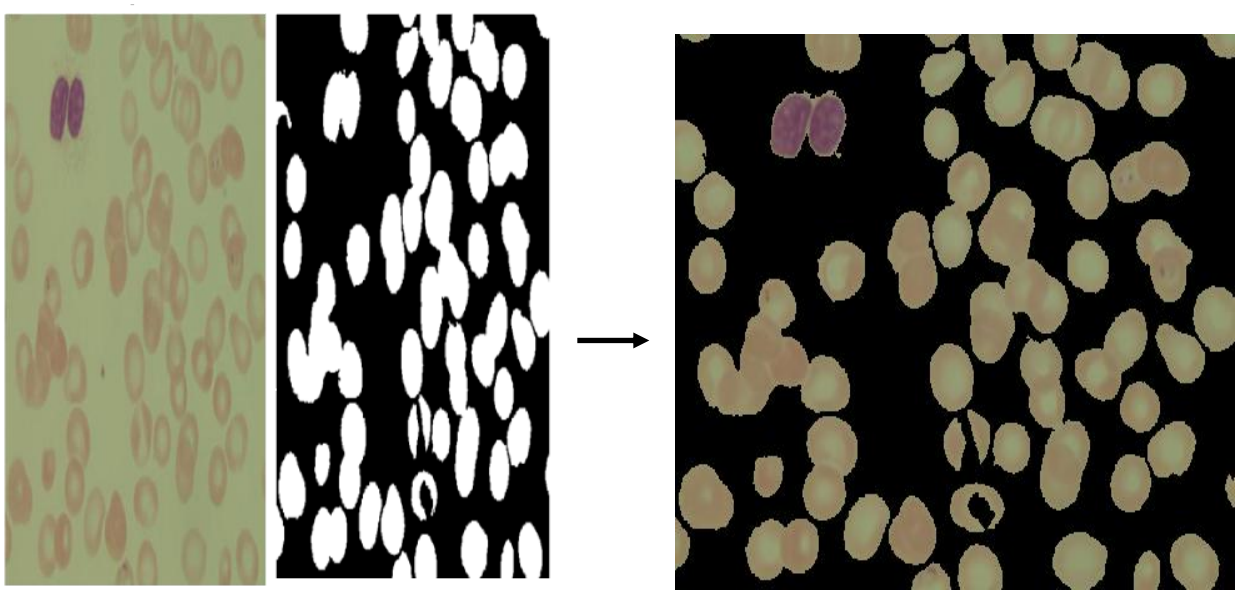


Figure 3.16: Segmented image with background set to 'zero' pixel values

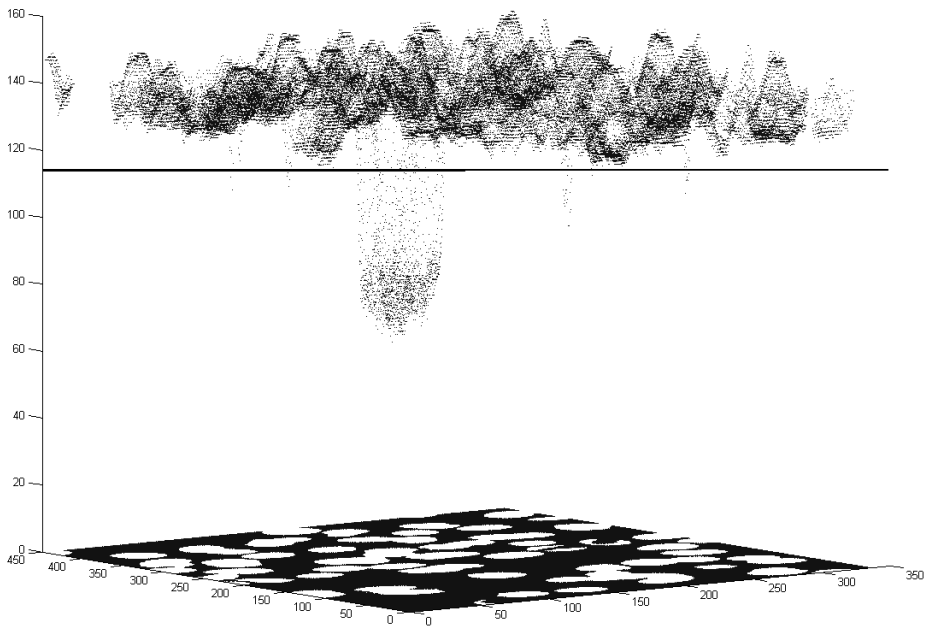


Figure 3.17: Pixel intensity plot with 'background pixel intensity' values set to zero

The parasite cluster as observed in Figures 3.15 and 3.17 is a weak cluster or in other words fewer pixels are stained. The small clusters (stained pixels) buried under a large population are clearly overwhelmed. Therefore the small cluster information is lost during the process of clustering and hardly any of the small clusters appear as separate cluster(s). The solution to this problem is obtained by imparting higher weights to the small cluster(s) compared to the larger clusters. In this way the small cluster information is retained by making small cluster comparable to big clusters. The new modified form of KMC as designed with the modifications discussed above is defined as *probabilistic k-means clustering* (PkMC). The discrete probability function (DPF) for PkMC was designed using the binomial theorem as described below. The binomial theorem facilitates the application of dynamic choice of higher weights to small cluster(s).

Binomial theorem:

The mathematical expression for the binomial theorem can be expressed as:

$$(x+y)^n = {}^n\mathbb{C}_0 x^n + {}^n\mathbb{C}_1 x^{n-1}y + \cdots + {}^n\mathbb{C}_k x^k y^{n-k} + \ldots + {}^n\mathbb{C}_p x^p y^{n-p} + \ldots + {}^n\mathbb{C}_{n-1} x y^{n-1} + {}^n\mathbb{C}_n y^n \quad (2)$$

Where:

$x, y \in \mathbb{R}^+$, $n \in \mathbb{N}$ and

$$ {}^n\mathbb{C}_k = \frac{n!}{k!(n-k)!} \quad (3)$$

The expression can be explained as follows: we have a total of 'n' clusters and the 'k^{th}' and 'p^{th}' clusters are to have the maximum weight. Since we only have one cluster (stained pixels) with the maximum weight, we can choose 'y' as unity, leaving variable 'x' free for maximization as represented in equation (4). The discrete probability distributions (DPF) for different cases are described below with the maximization equations.

$$(x+1)^n = {}^n\mathbb{C}_0 x^n + {}^n\mathbb{C}_1 x^{n-1} + \cdots + {}^n\mathbb{C}_k x^k + \ldots + {}^n\mathbb{C}_{n-1} x + {}^n\mathbb{C}_n \quad (4)$$

Case 1: When 'n' is even and for a given 'k^{th}' cluster p(k) can be designed from equation (4) is explained below;

$$ {}^n\mathbb{C}_1 x^{n-1} + \cdots + {}^n\mathbb{C}_k x^k + \ldots + {}^n\mathbb{C}_{n-1} x + {}^n\mathbb{C}_n = (x+1)^n - x^n \quad (5)$$

$$\Rightarrow \frac{{}^n\mathbb{C}_1 x^{n-1} + \cdots + {}^n\mathbb{C}_k x^k + \ldots + {}^n\mathbb{C}_{n-1} x + {}^n\mathbb{C}_n}{(x+1)^n - x^n} = 1 \quad (6)$$

$$\Rightarrow p(k) = \frac{{}^n\mathbb{C}_k x^{n-k}}{(x+1)^n - x^n} \quad (7)$$

Equation (7), shows that the probability of k^{th} cluster depends on x, k and n.

$$\lim_{x\to 0^+}\left(\frac{{}^{n}\mathbb{C}_{k}x^{n-k}}{(x+1)^{n}-x^{n}}\right)=0$$

$$\text{And } \lim_{x\to \infty}\left(\frac{{}^{n}\mathbb{C}_{k}x^{n-k}}{(x+1)^{n}-x^{n}}\right)=0$$

Mathematically, we have shown that p(k) will have a global maximum value for a particular value of x for a given set of k and n. The advantage of using this form of probability function with given n and k is that we can solve for x, such that p(k) is maximum.

The value of x for maximising p(k) can be solved by equating the derivative of p(k) with respect to 'x' to zero. The equation derived after equating the derivative to zero can be easily solved for a positive root of 'x' under the constraint that 0 < p(k) < 1:

$$n(1+x)^{n-1}-k(1+x)^{n}+kx^{n}=0 \tag{8}$$

Case 2: When 'n' is even and for a given 'kth' cluster p(k) can be designed from equation (4) as explained below for two separate sub cases:

Sub case 1: $1 \le k \le (n-1)/2$

$$p(k)=\frac{\left({}^{n}\mathbb{C}_{k-1}x^{n-k+1}\right)}{(1+x)^{n}-{}^{n}\mathbb{C}_{\frac{n-1}{2}}x^{\frac{n+1}{2}}} \tag{9}$$

Similar to the maximization scheme described earlier, the value of 'x' can be solved by equating the derivative of p(k) with respect to 'x' to zero. The equation derived after equating derivative to zero can be easily solved for positive of 'x' under the constraint 0 < p(k) < 1:

$$n(1+x)^{n-1}+(1-k)(1+x)^{n}+\left(\frac{2k-n-1}{2}\right){}^{n}\mathbb{C}_{\frac{n-1}{2}}x^{\frac{n+1}{2}}=0 \tag{10}$$

Sub case 2: $(n-1)/2 < k \le n$

$$p(k) = \frac{({}^{n}\mathbb{C}_{k} x^{n-k})}{(1+x)^{n} - {}^{n}\mathbb{C}_{\frac{n-1}{2}} x^{\frac{n+1}{2}}} \tag{11}$$

$$n(1+x)^{n-1} - k(1+x)^{n} + (\frac{2k-n+1}{2})^{n}\mathbb{C}_{\frac{n-1}{2}} x^{\frac{n+1}{2}} = 0 \tag{12}$$

In our case, ideally we should have three clusters (n = 3 and k = 2) to represent the background, RBC and parasite, thus the solution of equation (12) is x = 0.7937 and the corresponding DPF is shown in Figure 3.18.

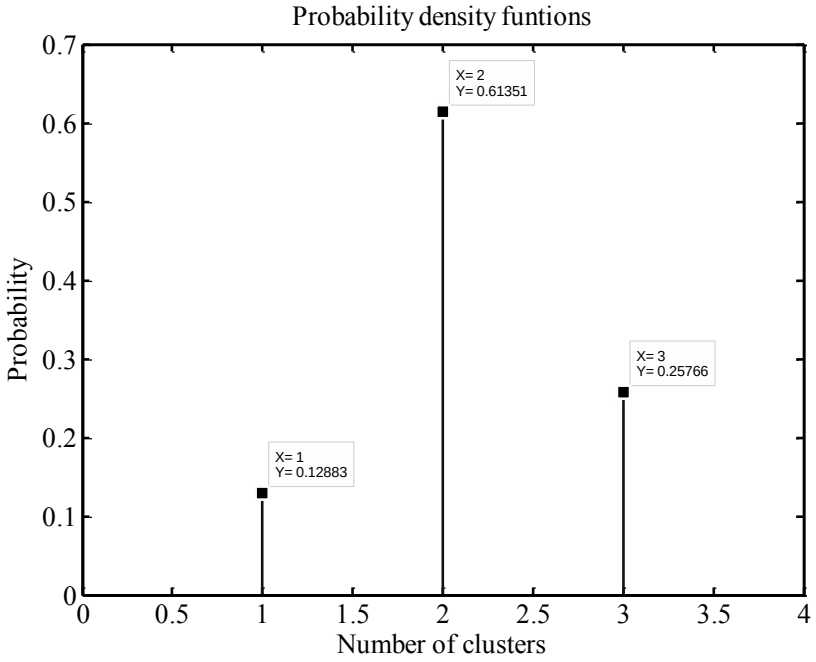


Figure 3.18: Probability density function for 3 clusters

Suppose we have some artefacts (such as the central pallor) which could be assigned as cluster. In this case our design parameters will be: number of clusters are 4 (n = 4), and the parasite would be represented by the third cluster (k = 3). The maximization scheme for this case yields, x = 1.2 and the corresponding DPF is as shown in Figure 3.19.

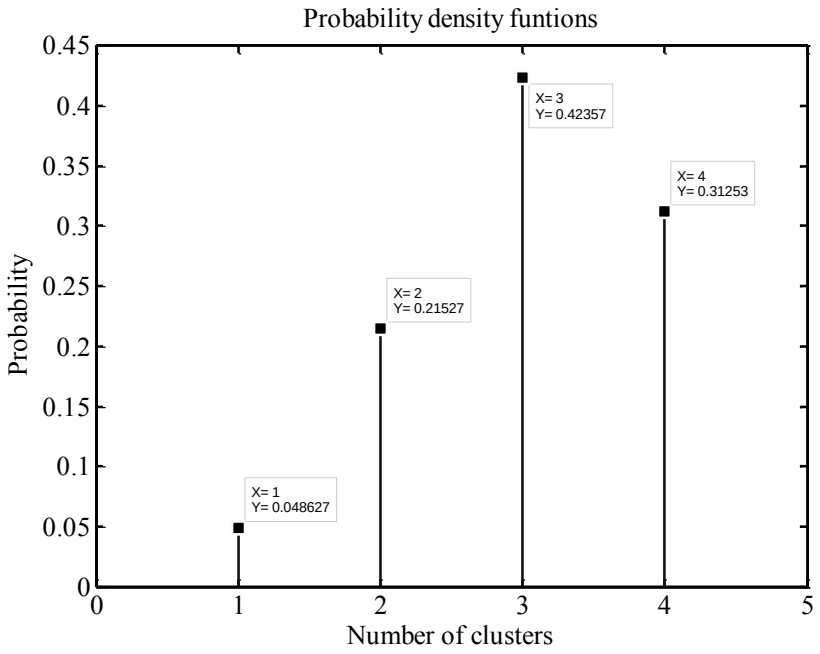


Figure 3.19: Probability density function for 4 clusters

The discrete probability density function designed as described above was used to impart weights to the clusters. The stained pixels were given higher weights and lower weights were assigned to other clusters so as to retain parasite information buried in a large population.

The Probabilistic (PkMC) labels the clusters as 1) the first cluster being for background, 2) the second cluster of stained cluster (i.e. parasites and arte facts) and 3) the third cluster of RBC. The PkMC also gives the average intensity of each cluster and designates this as the cluster centroids. The last part of the algorithm is concerned with how to separate artefacts from the potential parasites. The artefacts can be visualized as stained WBC, platelets, dirt or dye crystals.

3.4.4 Potential parasite confirmation:

The features of parasite need to be understood to separate them from artefacts. An intelligent separation scheme can be designed by exploiting the features such as: parasites are embedded with in the RBC and artefacts are more or less homogeneous objects. The method to separate parasites from the artefacts is described below by an

illustrative example. Figure 3.20a represents a sample image after background intensity set to zero and with a parasite in one of the RBCs. Figure 3.20b is a binary image of Figure 3.20a. Now each component in Figure 3.20a will have an average intensity and clearly the average intensity of a platelet or WBC will be near the stained pixel intensity because of their homogeneous nature. An intensity based threshold can be used to separate the platelets and WBCs. Recall that the PkMC method returns 3 clusters of a) background, b) stained pixels and c) RBC as shown in Figure 3.21. Apart from clusters labels PkMC also reports the average intensity of each cluster as three centroids. The comparison of the average intensity of each component from Figure 3.20a to the centroid for stained pixel helps to distinguish between platelet or WBC or any other artefact. This ensures that stained pixels which are parasites are retained and artefacts are removed as shown in Figure 3.22. Figure 3.22b represents a binary mask of stained pixels which happens to be a parasite; this binary mask is superimposed over the original image to give the actual parasite location and parasite count.

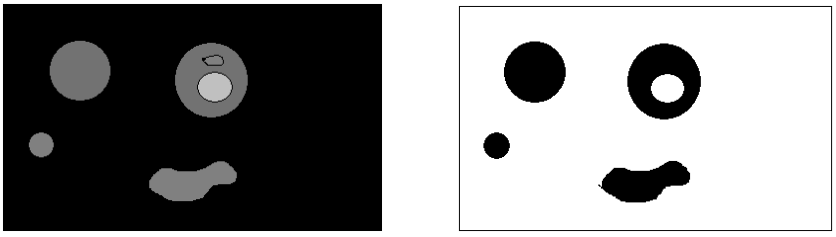


Figure 3.20: a) Gray scale image with background pixels set to zero (step 2) b) Binary image after boundary detection (step 5)

Figure 3.21: Ideally PkMC generates three clusters a) Background, b) Stained pixels (parasites, platelets and other artefacts) and c) RBC

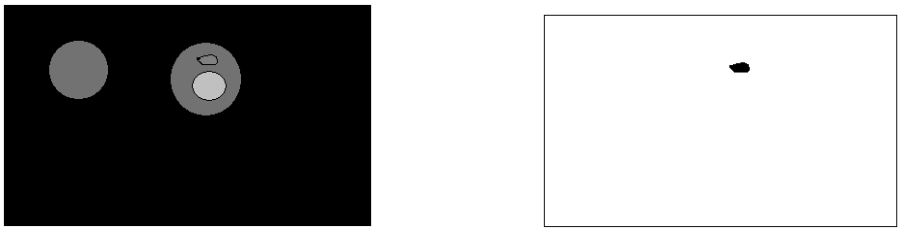
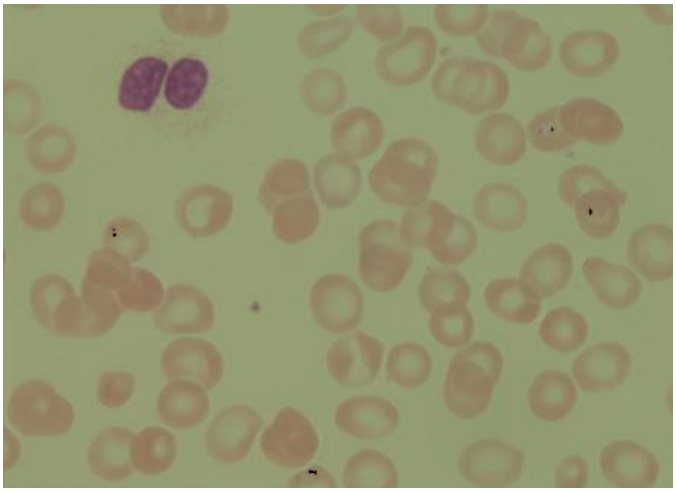
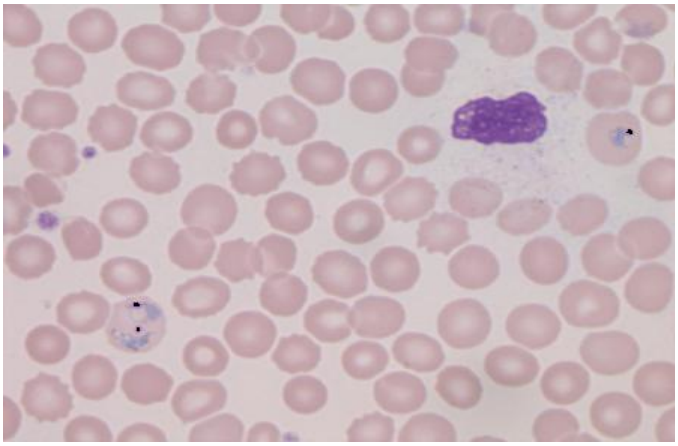


Figure 3.22: After implementing threshold check a) Only RBCs are retained i.e. platelet and artefact are removed to finally yield b) Binary mask of Parasite

The parasite detection method was implemented over the 10 potential cases and 10 blind cases. The detection of parasites in two examples is represented in Figure 3.23a and 3.23b are marked in black. Figure 3.23 shows the importance of unsupervised algorithm as both the images are taken under different condition. Also the platelets and WBC in the images are not recognized as parasites.



a) Parasites marked in black, Total infected RBC = 4



b) Parasites marked in black, Total infected RBC = 2

Figure 3.23: Parasites marked image

3.5 Results and discussion

The proposed method was tested on larger variety of images with Malarial parasites at different life cycle to check for the sensitivity and specificity of the algorithm. The performance of the method was tested over cases provided by Alberta Health Services for 10 potential cases of Malaria and 10 blind control cases to cross-validate the method. The measure of performance and accuracy of the method was determined under two parameters: sensitivity and positive predictive value (PPV) [9]. Sensitivity is defined as ability of the algorithm to correctly detect parasite(s). PPV is defined as the precision rate of the algorithm. The values for sensitivity and PPV are expressed in terms of true positives (TP), false positive (FP) and false negative (FN) as defined in expressions (10) and (11):

$$Sensitivity = \frac{TP}{TP + FN} \tag{10}$$

$$PPV = \frac{TP}{TP + FP} \tag{11}$$

The results of all the cases are summarized below, the values of false positive, false negative and true positive are reported in Table 3.

Table 3 gives a good insight of the automated diagnosis process and the decision making by the algorithm. The large number of false positives can be attributed to the presence of a higher number of artefacts and poor image quality. The benefit of the unsupervised method proposed here can be seen with minimal or zero cases of false negatives.

Total cases	20
True positives	15
False positives	5
False negatives	0
Sensitivity %	100
Positive Predictive Value %	75

Table 3.2: Results for detecting and confirming parasites

Apart, from the combined results of 20 cases a detailed analysis for one random case is also represented. The results from each of the 10 images for random case are reported in Table 4. Table 4 gives a comparative study between results from the proposed algorithm and results from the pathologist report. This study confirms the sensitivity and accuracy of the proposed method. The small errors in the infected RBC count were due to high platelet count. The results for all the cases are reported separately at the end of this chapter.

	Algorithm result		Pathologist's result	
	RBC count	Infected RBC count	RBC count	Infected RBC count
M7-1	67	0	66	0
M7-2	74	2	74	2
M7-3	64	3	64	3
M7-4	75	0	73	0
M7-5	64	1	64	1
M7-6	76	2	76	1
M7-7	75	3	75	3
M7-8	72	5	71	1
M7-9	64	1	66	1
M7-10	74	1	75	1

Table 3.3: Results for patient M7, reported parasitemia = 2.55%

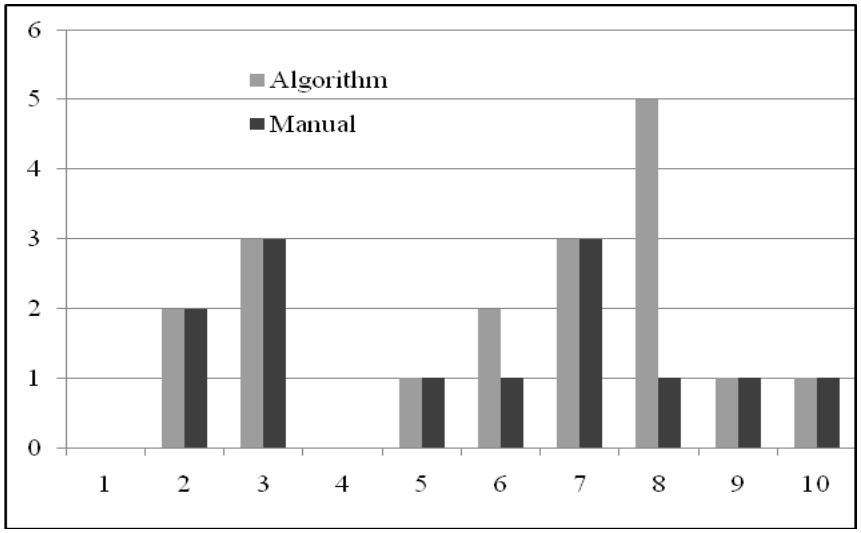


Figure 3.24: Comparison of infected RBC count

Clearly the parasitemia level is a crucial factor to report and dependent over the number of images observed. This is a fundamental limitation of microscopy, that is the ability to detect Malaria depends on the number of microscopic fields observed [11]. The automated technique presented here is successful in reporting parasitemia level of 0.1% or less. This can be understood as follows: the proposed algorithm can even capture a single infected RBC out of 1000 observed RBCs.

The advantage of the method presented in this chapter is to simply report parasitemia. The main benefit is that it has significantly reduced the effective time of diagnosis when combined with an image acquisition system, and using this method it is possible to examine entire slide. Secondly, the unsupervised nature of the method will minimize the human reliance and will bring down the effective cost of diagnosis. In addition to this, the method not only focuses on detecting parasites, it also reports the size distribution of RBCs which can contribute in detecting other diseases related to size of RBC and can be used an offline RBC counter. Lastly, the method is reliable, automated or un-supervised and gives consistent results.

3.6 Concluding remarks

This chapter provides an automated, robust and unsupervised screening method for Malaria parasites. Existing diagnosis methods in the literature depend extensively on skilled practitioners and require training, that is, they are supervised or non-automated. Compared to other diagnostic techniques, it avoids problems associated with rapid methods, such as being species specific and high per test cost. In contrast to this the method proposed here is completely automated, following the slide making and initial digital imaging of slides. The method was specifically developed to work on low resolution images and thereby removes dependency on high resolution microscopy which boosts the practicality of the algorithm. Combined with an image acquisition system will allow entire slide to be examined minimizing human reliance. The necessity of electricity is principal obstacle in using such kind of system.

The algorithm was not only developed to detect parasite but it also determines the RBC size distribution which may be of use in the diagnosis of other diseases.

3.7 Complete case study for potential and control cases

3.7.1 Potential cases:

M1: Parasitemia = 0.64%

	Algorithm count	
M1-1	82	0
M1-2	93	0
M1-3	92	0
M1-4	91	0
M1-5	115	0
M1-6	101	0
M1-7	91	0
M1-8	83	0
M1-9	93	6
M1-10	91	0

M3: Parasitemia = 0.73%

	Algorithm count	
M3-1	85	0
M3-2	87	0
M3-3	83	0
M3-4	87	1
M3-5	71	0
M3-6	87	1
M3-7	78	0
M3-8	67	0
M3-9	96	0
M3-10	81	4

M4: Parasitemia = 2.69%

	Algorithm count	
M4-1	54	3
M4-2	64	1
M4-3	59	1
M3-4	59	2
M4-5	61	1
M4-6	55	0
M4-7	71	3
M4-8	61	5
M4-9	51	0
M4-10	59	0

M5: Parasitemia = 0.59%

	Algorithm count	
M5-1	63	4
M5-2	65	0
M3-3	70	0
M5-4	78	0
M5-5	56	0
M5-6	77	0
M5-7	68	0
M5-8	68	0
M5-9	65	0
M5-10	59	0

M6: Parasitemia = 0%

	Algorithm count	
M6-1	117	0
M6-2	88	0
M6-3	98	0
M6-4	98	0
M6-5	101	0
M6-6	109	0
M6-7	98	0
M6-8	101	0
M6-9	110	0
M6-10	104	0

M7: Parasitemia = 2.55%

	Algorithm count	
M7-1	67	0
M7-2	74	2
M7-3	64	3
M7-4	75	0
M7-5	64	1
M7-6	76	2
M7-7	75	3
M7-8	72	5
M7-9	64	1
M7-10	74	1

M8: Parasitemia = 3.11%

	Algorithm count	
M8-1	82	0
M8-2	84	0
M8-3	87	0
M8-4	84	1
M8-5	90	0
M8-6	81	1
M8-7	83	0
M8-8	84	2
M8-9	74	0
M8-10	85	2

M9: Parasitemia = 1.06%

	Algorithm count	
M9-1	101	0
M9-2	104	0
M9-3	103	0
M9-4	100	3
M9-5	100	0
M9-6	99	0
M9-7	92	0
M9-8	84	5
M9-9	74	0
M9-10	85	2

M10: Parasitemia = 4.47%

	Algorithm count	
M10-1	79	0
M10-2	91	0
M10-3	104	7
M10-4	94	6
M10-5	91	0
M10-6	85	7
M10-7	85	8
M10-8	75	8
M10-9	90	0
M10-10	78	3

M11: Parasitemia = 0.488%

	Algorithm count	
M11-1	67	1
M11-2	78	0
M11-3	73	0
M11-4	82	2
M11-5	77	0
M11-6	85	0
M11-7	87	0
M11-8	93	0
M11-9	83	1
M11-10	93	0

3.7.2 Control cases:

C1: Parasitemia = 1.26%

	Algorithm count	
C1-1	80	0
C1-2	80	0
C1-3	80	2
C1-4	76	0
C1-5	76	0
C1-6	74	0
C1-7	66	1
C1-8	78	0
C1-9	98	7
C1-10	85	0

C2: Parasitemia = 0%

	Algorithm count	
C2-1	72	0
C2-2	64	0
C2-3	72	0
C2-4	71	0
C2-5	65	0
C2-6	93	0
C2-7	75	0
C2-8	79	0
C2-9	66	0
C2-10	64	0

C3: Parasitemia = 0%

	Algorithm count	
C3-1	87	0
C3-2	82	0
C3-3	84	0
C3-4	90	0
C3-5	78	0
C3-6	90	0
C3-7	87	0
C3-8	82	0
C3-9	73	0
C3-10	86	0

C4: Parasitemia = 1.65%

	Algorithm count	
C4-1	65	0
C4-2	64	0
C4-3	64	0
C4-4	54	0
C4-5	61	0
C4-6	53	0
C4-7	61	10
C4-8	66	0
C4-9	58	2
C4-10	64	0

C5: Parasitemia = 0%

	Algorithm count	
C5-1	68	0
C5-2	60	0
C5-3	61	0
C5-4	58	0
C5-5	58	0
C5-6	61	0
C5-7	50	0
C5-8	68	0
C5-9	-	-
C5-10	63	0

C6: Parasitemia = 0.99%

	Algorithm count	
C6-1	68	3
C6-2	64	2
C6-3	71	0
C6-4	67	0
C6-5	77	0
C6-6	67	0
C6-7	75	0
C6-8	71	1
C6-9	68	0
C6-10	76	1

C7: Parasitemia = 3.00%

	Algorithm count	
C7-1	74	0
C7-2	70	1
C7-3	84	3
C7-4	73	0
C7-5	57	7
C7-6	81	2
C7-7	62	6
C7-8	61	0
C7-9	66	2
C7-10	55	0

C8: Parasitemia = 3.48%

	Algorithm count	
C8-1	54	1
C8-2	52	0
C8-3	58	3
C8-4	56	0
C8-5	64	1
C8-6	59	2
C8-7	66	5
C8-8	59	4
C8-9	61	5
C8-10	74	0

C9: Parasitemia = 0%

	Algorithm count	
C9-1	60	0
C9-2	37	0
C9-3	46	0
C9-4	47	0
C9-5	31	0
C9-6	43	0
C9-7	64	0
C9-8	65	0
C9-9	53	0
C9-10	44	0

C10: Parasitemia = 0%

	Algorithm count	
C10-1	73	0
C10-2	81	0
C10-3	73	0
C10-4	70	0
C10-5	31	0
C10-6	43	0
C10-7	64	0
C10-8	65	0
C10-9	53	0
C10-10	44	0

Chapter 4

Concluding Remarks & Future Work

4.1 Concluding remarks

This thesis has successfully demonstrated the significance of data based methods for abnormality detection and quantification. The work explored two different streams of abnormality or anomaly detection and has demonstrated that data based diagnosis can indeed be effective in various fields of science. The non-intrusive and non-destructive nature of these methods has a clear advantage over other methods. This thesis has provided a 'complete solution' for the digital image based diagnosis of Malaria. Also, it has closely looked into the bi-coherence based method to detect and quantify valve problems. The suggested revisions and improvements to the existing algorithm have demonstrated that the solution suggested is even more broadly applicable to all types of valve problems and not simply valve stiction.

The data driven methods have proven to be most successful to address the issues of valve faults in controlled loops. The bi-coherence based method has proven to be very successful in stiction detection. However, the method restricts itself only to the detection and diagnosis of valve stiction. If we closely observe industrial data we can see that stiction is only one of many possible valve faults. Secondly, the existing bi-coherence based method tends to generate a few cases of missed and false alarms. To minimize this issue, the bi-coherence method has been revised wherein each step was examined in detail. A revised logic for the algorithm was proposed and is shown in Figure 2.21. The new logic for non-linearity index (NLI) check was successfully able to capture intermittent valve problems. The revision of this 'NLI check' has served to reduce the number of missed alarms. The revised 'set point (SP) activity' check combined with efficient non-linearity check has made the 'SP activity check' very effective leading to significant reduction in the number of false alarms. Lastly, the revision in NLI

reporting as: 'percentage of data windows with high NLI' has given operator access to more detailed information of the data set, including confidence margins in the diagnosis.

In chapter 2, we have shown the efficacy of these suggested changes through two industrial case studies. The comparative study in section 2.5 shows significant improvement in the overall data based diagnosis of valve controlled loops.

Chapter 3 explored the field of image analysis and presented a 'complete solution' for image based Malaria diagnosis. Image based diagnosis is an open area of study and research and has gained immense popularity in the medical field. This chapter has elucidated the problems concerning Malaria and indicated the potential impact of such automated methods. The need for an automated diagnosis is clearly evident from the number of casualties and from the volume of infections caused by lack of cheap and speedy diagnosis. The chapter provides an automated, robust, unsupervised and practical method for malarial parasite detection in thin smeared blood images. The schematic of the method proposed is shown Figure 3.4. The proposed method is theoretically simple and robust.

The lack of high resolution microscopes in developing countries was carefully considered by developing a method to work with low resolution images. Low resolution images made image segmentation challenging with conventional edge detection techniques. To have effective level of image segmentation the method of energy minimization based boundary detection has proved its efficacy over conventional edge detection techniques. The novel idea for 'hole-filling' algorithm combined with circle detection method has closely replicated the method of RBC enumeration as done in manual microscopy. Apart from RBC enumeration, this method can be easily tuned in for automatic detection of circular objects embedded in an image.

A substantial amount of research effort has been dedicated towards understanding clustering based image segmentation. The novel concept of binomial based

probability distribution was specially tailored to capture the information of barely visible parasite(s). The introduction of probabilistic clustering (PkMC) has proven to be an excellent approach to capture small cluster(s) buried in a large population. The PkMC method is sensitive enough to capture a single parasite in an image. When combined, the RBC enumeration with parasite detection method end up as an unsupervised algorithm that needs no initial training and it certainly has advantages over other methods in the literature. Apart from diagnosis of malaria, the algorithm can easily be tuned for processing other stream of images.

4.2 Future work

4.2.1 Coherency threshold

Chapter 2 presented the importance of time lagged correlation between SP and PV to avoid false alarms. But determining the temporal time lag correlation in noisy data is not an easy task to perform. The issues concerning temporal time-lagged correlation can be avoided by replacing it with spectral correlation. The idea proposed is to use coherency metric to capture spectral correlation. Future researchers can develop an analytical threshold from coherency metric to capture time-lag adjusted correlation.

4.2.2 Oscillation detection

Oscillation detection is an important step in overall detection and quantification of valve problems as discussed in chapter 2. The existing oscillation detection algorithm is dependent on the condition of stationarity of the signal. Oscillation detection algorithm tends to give undesired results for industrial cases where data is not stationary. This opens a good opportunity for research to explore areas where data stationarity in not mandatory.

4.2.3 Image acquisition system

Chapter 3 demonstrates the importance of computer vision based diagnosis of Malaria. Combined with an automated digital microscope, this would allow an entire slide to be examined. High electricity cost and requirement of computers

are obstacles to use this system in field particularly in developing countries. This gives a good opportunity to develop a complete system which is capable of image acquisition with simultaneous processing ability. The device can be very handy particularly in remote areas where access to electricity and computers is limited.

4.2.4 Enhance reliability of automated diagnosis

Medical imaging has attracted significant attention for automatic diagnosis. Somehow, the penetration of automated diagnosis is not easy in medical fields. The 'reliability' of diagnostic scheme is an important issue for automated method in the medical field. 'Reliability' depends on the sensitivity, specificity and practicality of a method. The reliability issue can be addressed by testing diagnostic schemes for large sets of images and compares them with the manual microscopy method. This motivates the need to invest a substantial amount of effort to collect images from different laboratories and produce extensive case study of comparative results.

References

1. Choudhury M A A S, Shah S L and Thornhill N F, Diagnosis of Process Nonlinearities and valve stiction Data Driven Approaches. Springer-Verlag. Germany.
2. Chan TF, LA Vese: Active Contours Without Edges, IEEE Transaction on Image Processing, Vol. 10, No.2 Feb. 2001.
3. Gonzalez RC, Woods RE, Eddins SL: Digital Image Processing using MATLAB, Prentice Hall 2001.
4. Gonzalez RC, Woods RE: Digital Image Processing, 2nd Edition: Prentice Hall 2001.
5. Hanscheid, T., 1999. Diagnosis of malaria: a review of alternatives to conventional microscopy. Clin. Lab. Haematol. 21.
6. http://www.time.com/time/health/article/0,8599,1842102,00.html
7. Iancu DA: Eye detection using variants of the Hough Transform: Final project in CS475 Computational Vision and Biological perception 2004.
8. Jelali M and B. Huang (2009), Detection and Diagnosis of Stiction in Control Loops: State of the Art and Advanced Methods, Springer Verlag, Germany.
9. Jampana PV: Computer Vision based Sensors for Chemical Process. PhD thesis, University of Alberta, Department of Chemical and Materials Engineering, 2010.
10. Johnson RA, Wichern DW: Applied Multivariate Statistical Analysis 5th edition, Pearson education 2009.
11. Oaks SCJ, Mitchell VS, Pearson GW, Carpenter CCJ: Malaria: Obstacles and opportunities, National Academy Press, Washington, DC. A report of the committee for the study on malaria prevention and control Status review and alternative strategies.

12. Ross NE, Pritchard CJ, Rubin DM, Duse: Automated image processing method for the diagnosis and classification of malaria on thin blood smears. Med boil Comput 2006 44: 427-436.
13. Rao KNRM, Dempster AG, Jarra B, Khan S: Automatic scanning of malaria infected blood slide images using mathematical morphology. Proc IEE Seminar Medical Applications of Signal Processing, London, UK 2002.
14. Rao KNRM: Application of mathematical morphology to biomedical image processing. *PhD thesis*. University of Westminster; 2004.
15. Shah S, The Fever: How malaria has ruled humankind for 500,000 years, Sarah Crichton Books.
16. Sio SW, Sun W, Kumar S, Bin WZ, Tan SS, Ong SH, Kikuchi H, Oshima Y, Tan KS: MalariaCount: An image analysis-based program for the accurate determination of parasitemia. Journal of Microbiological Methods Volume 68, Issue 1, January 2007, Pages 11-18
17. Tek FB, Dempster AG, Kale I: Parasite detection and identification for automated thin blood film malaria diagnosis. Computer Vision and Image Understanding 114 (2010) 21-32
18. Tek FB, Dempster AG, Kale I: Computer vision for microscopy diagnosis of malaria. Malaria Journal 2009,8:153.
19. World Malaria Report 2008. In Tech rep. World Health Organization, Geneva; 2008
20. WHO: Basic Microscopy Part I. Learner's Guide: World Health Organization; 1991