August 2014

Human Metaphase Chromosome Analysis using Image Processing

Akila M.S Subasinghe Arachchige
The University of Western Ontario

Supervisor
Dr. Jagath Samarabandu
The University of Western Ontario

Graduate Program in Electrical and Computer Engineering

A thesis submitted in partial fulfillment of the requirements for the degree in Doctor of Philosophy

Recommended Citation

Subasinghe Arachchige, Akila M.S, "Human Metaphase Chromosome Analysis using Image Processing" (2014). *Electronic Thesis and Dissertation Repository*. 2178.
https://ir.lib.uwo.ca/etd/2178

Human Metaphase Chromosome Analysis using Image Processing

(Thesis format: Monograph)

by

Akila Subasinghe Arachchige

Graduate Program
in
Engineering Science
Electrical and Computer Engineering

A thesis submitted in partial fulfillment
of the requirements for the degree of
Doctorate of Philosophy

School of Graduate and Postdoctoral Studies
The University of Western Ontario
London, Ontario, Canada

Abstract

Development of an effective human metaphase chromosome analysis algorithm can optimize expert time usage by increasing the efficiency of many clinical diagnosis processes. Although many methods exist in the literature, they are only applicable for limited morphological variations and are specific to the staining method used during cell preparation. They are also highly influenced by irregular chromosome boundaries as well as the presence of artifacts such as premature sister chromatid separation.

Therefore an algorithm is proposed in this research which can operate with any morphological variation of the chromosome across images from multiple staining methods. The proposed algorithm is capable of calculating the segmentation outline, the centerline (which gives the chromosome length), partitioning of the telomere regions and the centromere location of a given chromosome. The algorithm also detects and corrects for the sister chromatid separation artifact in metaphase cell images. A measure termed the Candidate Based Centromere Confidence (CBCC) is proposed to accompany each centromere detection result of the proposed method, giving an indication of the confidence the algorithm has on a given localization.

The proposed method was first tested for the ability of calculating an accurate width profile against a centerline based method [1] using 226 chromosomes. A statistical analysis of the centromere detection error values proved that the proposed method can accurately locate centromere locations with statistical significance. Furthermore, the proposed method performed more consistently across different staining methods in comparison to the centerline based approach. When tested with a larger data set of 1400 chromosomes collected from a set of DAPI (4',6-diamidino-2-phenylindole) and Giemsa stained cell images, the proposed candidate based centromere detection algorithm was able to accurately localize 1220 centromere locations yielding a detection accuracy of 87%.

Acknowledgements

First and foremost, I would like to thank my main supervisor, Dr. Jagath Samarabandu for all the invaluable guidance and advice given to me throughout my PhD, both academically and personally. He has truly been a big influence and a role model for my academic career as well as for my personal life. I would like to thank my co-supervisor Dr. Peter Rogan as well as Dr. Joan Knoll for all their guidance and knowledge shared with me during my research. A great deal of appreciation needs to be shown to my course instructors, Dr. Ladak, Dr. Olga Veksler, Dr. Yuri Boykov and Dr. John Barron for their innovative and attractive ways of teaching and motivating my work. I would also like the thank Dr. Quazi Rahman for guiding me through my TA duties during this period.

Mr. Wahab Khan, Ms. Natasha Caminsky, Ms. Heather Tarnowski, Dr. Ruth Wilkins and Ms. Farrah Flegal were instrumental in collecting data required for the research and requires special acknowledgement. Contributions of my research colleagues Mr. Yanxin Li and Mr. Asanka Wickramasinghe are also mentioning with a thankful note for their efforts in converting the proposed work to C++ code. I would also like to thank Dr. Jacques for the valuable discussion we have had both on research matters as well as on photography.

I am extremely thankful for the financial support provided by the Western Innovation Fund (University of Western Ontario), Natural Sciences and Engineering Research Council of Canada and the DART-DOSE CMCR (5U01AI091173-02 from the NIAID/US Public Health Service).

I would also like to thank my mother and my late father, who passed away during the first year of my masters research. I would not be here if not for the courage, guidance and support given by my parents who idolize and cherish all my achievements. The small Sri Lankan community in London needs to be acknowledged for caring for me throughout these past two years. I would also like to acknowledge my 2 year old son who brings a lot of love and joy into my life. Last, but definitely not least, I would like to thank my beloved wife whom I draw a lot of energy and courage from in times of need. She is also my first proof reader, who in many ways was actively involved in finishing this dissertation.

Akila Mike,
June - 2014,
London ON.

Table of Contents

Abstract . . . . **ii**

Acknowledgements . . . . **iii**

List of tables . . . . **vi**

List of figures . . . . **viii**

Acronyms and Symbols . . . . **xii**

1 Introduction **1**
1.1 Contributions 5
1.2 Thesis organization 8

2 Background **9**
2.1 Introduction to human chromosomes 9
2.1.1 The centromere 10
2.1.2 The telomere 14
2.1.3 The centromere index (CI) 14
2.2 Review of existing algorithms 16
2.2.1 Segmentation methods 16
2.2.2 Centromere detection methods 19
2.3 Background methods 22

3 Proposed algorithm **24**
3.1 The data set 25
3.2 Pre processing and segmentation 28
3.3 Finding the centerline 33
3.3.1 Centerline based Centromere identification 40
3.4 Contour partitioning & correcting for sister chromatid separation . . 41
3.4.1 Contour partitioning for isolating the telomere 43
3.4.2 Shape information extraction 49
3.5 Laplacian based thickness measurement 53
3.5.1 Intensity integrated Laplacian based thickness measurement . 54
3.6 Candidate point generation & metaphase centromere detection 61
3.6.1 Principal component analysis (PCA) of features 65
3.6.2 Candidate based centromere confidence (CBCC) 66

4 Results . . . 73
4.1 Preliminary testing . . . 73
4.1.1 Statistical analysis . . . 75
4.2 Candidate based method performance . . . 84

5 Conclusions & Future work . . . 91
5.1 Summary of the proposed method . . . 91
5.2 Conclusive remarks . . . 93
5.3 Future work . . . 94

Appendices . . . 96

A Background methods . . . 97
A.1 GVF snakes . . . 97
A.1.1 Parametric snakes or active contours . . . 97
A.1.2 Gradient vector flow as an external energy . . . 101
A.1.3 GVF snakes vs DT snakes . . . 102
A.2 Discrete curve evolution (DCE) . . . 104
A.2.1 Definitions . . . 105
A.2.2 The relevance function . . . 108
A.2.3 Advantages & disadvantages of DCE . . . 110
A.3 Support Vector Machines (SVM) . . . 112
A.3.1 Deriving the SVM framework . . . 114

Bibliography . . . 116

Curriculum Vitae . . . 125

List of Tables

2.1 Typical range of centromere index (CI) for each human chromosome [2] 15

3.1 A kernel that represents the Laplacian equation on a digital image . . 57
3.2 The kernel that integrates intensity information into the Laplacian calculation for location (x, y) in a $3x3$ local neighborhood. 59
3.3 The percentage contribution of each feature to the variance of the whole data set in the descending order. 66
3.4 The mean and the standard deviation of the CBCC values in cases with accurate centromere detection as well as inaccurate centromere detection. The table also includes cases where non of the candidates were not found to be viable candidates (a subset of inaccurate centromere detection) for the centromere location. 67
3.5 The mean and the standard deviation of the percentile confidence values in cases with accurate centromere detection as well as inaccurate centromere detection. The table also includes cases where none of the candidates were not found to be viable candidates (a subset of inaccurate centromere detection) for the centromere location. 69

4.1 Breakdown of chromosome cell images and chromosomes based on the staining method and the sister chromatid separation (SC Sep.) 74
4.2 The Kolmogorov-Smirnova normality test results for the data set. . . 78
4.3 Descriptive values for the detection error data set when analyzed with proposed Laplacian based method (E_L) and Centerline based method (E_C) [1]. 78
4.4 Levene's test for testing equal variance of detection error within image groups (given in table 4.1) for each algorithm. 78
4.5 The Wilcoxon signed test rank analysis results. 79
4.6 The Wilcoxon signed test significance analysis results. 79
4.7 Descriptive values for the detection error for the proposed Laplacian method (E_L) and the centerline based method (E_C) with respect to the image groups stated in table 4.1. 80
4.8 ANOVA test for significance within image groups. 81
4.9 Games-Howell post-hoc test results for analyzing significance of variance between image groups where the groups Group 1,2,3 were D-NSC (DAPI without SC Sep.),D-WSC (DAPI with SC Sep.) and G-WSC (Giemsa with SC Sep.) respectively. 81

4.10 Non-parametric correlation test for significance (*Sig.*) using Spearman coefficient for detection error between the proposed algorithm and the centerline based method where $\rho_{(X,Y)}$ denotes the correlation between two random variabels. 83
4.11 Breakdown of chromosome cell images and chromosomes used for the larger data set based on the staining method and the sister chromatid separation (SC Sep.) 84
4.12 The detection accuracy values for chromosomes used for the larger data set based on the staining method and the sister chromatid separation (SC Sep.) 87
4.13 The results of the preliminary analysis in studying the feasibility of extending the proposed method for dicentric detection is presented by indicating the number of times different ranked candidates were able to encompass the second centromere. 90

List of Figures

2.1 An illustration of the structure of a chromosome in context of the cell (Source- http://commons.wikimedia.org/wiki/File:Chromosome.gif). . 11

2.2 A karyotype analysis end result with all 46 chromosomes organized according to their chromosome groups and types (Reproduced with permission from Dr. Joan Knoll and Dr. Peter Rogan). 12

2.3 The structural components of metaphase chromosomes of all three categories : the metacentric, sub-metacentric and acrocentric 13

2.4 The lengths used for calculating the centromere index of a given chromosome. 14

3.1 The flow diagram of the proposed method. 26

3.2 Shows two cell images with different staining methods. Figure 3.2(a) contains a cell image with DAPI staining while figure 3.2(b) is a Giemsa stained image. 27

3.3 The flow diagram of the preprocessing and segmentation stage of the proposed method. 29

3.4 The window-center intensity mapping scheme which was used to map a certain intensity range (defined by the window and the center) to the full range (intensity levels 0 - 255). 30

3.5 Depicts an example of the effect of intensity normalization on a DAPI stained image window. Figure 3.5(a) & (b) depicts the original image window and the corresponding histogram. Similarly the figures 3.5(c) & (d) depicts the intensity normalized image window and the corresponding histogram. 31

3.6 Depicts the resulting image window at different stages of the segmentation algorithm where figure 3.6 (a) gives the original window containing the chromosome prior to segmentation. Figures 3.6 (b) and (c) contain the threshold output and the extracted binary object. The GVF outcome is given in figure 3.6 (d) 34

3.7 Two chromosomes and their skeletonization [3] and morphological thinning results [4] showing some resulting spurious branches. These operations were performed on the binary object obtained through Otsu's method. Figure 3.7 (b) and (e) are the skeletonization results while figure 3.7 (c) and (f) are the morphological thinning results. 36

3.8 Comparison between standard skeleton with DCE based skeletal pruning results where (c) and (d) are the skeletons resulting from the DCE triangle and the DCE pentagon results. 38

3.9 The eight masks used for the hit and miss algorithm where '+1' and '-1' elements search for matching foreground and background pixels respectively, while the '0' ignores the value at that position. 39

3.10 Depicts the trellis structure created for calculating the width profile of a chromosome. 40

3.11 (a) and (b) depict two instances where the centerline based centromere detection was adversely affected by sister chromatid separation. . . . 41

3.12 The flow diagram of the correction for sister chromatid separation stage of the proposed method. 42

3.13 Depicts the effects on the chromosome morphology introduced by sister chromatid separation. (a) and (b) depict two straight chromosomes without and with sister chromatid separation respectively both with DAPI staining. 43

3.14 Figure demonstrates one possible combination for contour partitioning where the anchor point (red '+' sign) P_1^E is connected with the candidate point P_4^S while the other anchor point P_2^E is connected with candidate point P_1^S which captures the telomere regions. The blue line connects the set of points constituting the considered combination in this instance. 47

3.15 The flow diagram of the Intensity integrated Laplacian algorithm of the proposed method. 55

3.16 Depicts the uniform sampling of the width profile using the proposed method in figure 3.16 (c) & (d) as opposed to the trellis structure measurements through the centerline based approaches given by figure 3.16 (a) & (b). 69

3.17 Demonstrates the difference between the kernel of the proposed method in comparison to the standard Laplacian kernel. Figure 3.17(b) is an enlarged view of the 3x3 neighborhood of the pixel location marked by yellow on figure 3.17(a). Figure 3.17(c) & (d) represents the standard Laplacian and intensity biased Laplacian kernels calculated for the neighborhood of interest. 70

3.18 The steps of tracing the thickness (yellow stars) at one contour location of the chromosome where the arrows indicate the Laplacian vector field. The black square indicates the end point on the contour of the object. The final thickness value is calculated by getting the sum of all the lengths of these small steps. 70

3.19 An example where the contour C is split into two approximately symmetric segments C^1 and C^2. The red width trace line connects the points C_λ^1 and C_λ^2 of the two contour segments. 71

3.20 Shows the expected scenario for candidate based centromere detection where the blue square represents the optimal candidate while the other five candidates are given by the red squares in the feature space. . . . 71
3.21 A collection of possible cumulative distributive functions (of Gaussian distributions) where the (μ, σ) values are set at (0.5, 0.5) for figure 3.21(a) , at (0.5, 0.1) for figure 3.21(b), at (0.75, 0.5) for figure 3.21(b) and finally at (0.75, 0.1) for figure 3.21(d). 72

4.1 Demonstrates some sample results of the algorithm where the detected centromere location is depicted in red color circle against the centerline based approach [1] in blue color star while the ground truth centromere line is depicted in white. Figure 4.1 (a)&(b) are results of DAPI (4',6-Diamidino-2-Phenylindole) stained chromosomes while figure 4.1 (c)-(f) are results of Giemsa stained chromosomes. Figure 4.1(e) is an instance in which the proposed algorithm has outperformed the state of the centerline method significantly while figure 4.1 (f) is an instance in which the proposed algorithm has failed to yield the accurate centromere location due to high degree of sister chromatid separation. . . 76
4.2 Depicts an example of the correction for the sister chromatid separation artifact and the impact of that correction on the width profile measurement in the proposed method. The right hand side telomere region was corrected for sister chromatid separation in figure 4.2 (c) where the resulting uniform sampling of the width profile from the Laplacian based method is given in figure 4.2 (d) as opposed to the centerline based method in figure 4.2 (b). 77
4.3 The scatter plots for demonstrating the correlation between the two detection error distributions in which the 'x' axis is the detection error of the proposed Laplacian based method (E_L) and the 'y' axis is the centerline based method (E_C). 82
4.4 Demonstrates an example where 5 candidates were created for the chromosome in figure 4.4 (a) using the width profile in figure 4.4 (b). The figure 4.4 (c) shows the signed distance values for each candidate calculated from the separating hyperplane while the selected candidate is depicted in blue. 86
4.5 Demonstrates some sample results of the algorithm where the accurately detected centromere location (selected candidate) is depicted by a yellow dot while the segmented outline is drawn in blue. Figure 4.5 (a) is a result of DAPI stained chromosomes while figure 4.5 (b)-(f) are results of Giemsa stained chromosomes. These results reported CBCC measures of (a) 1.000, (b) 1.000, (c) 1.000, (d) 0.995, (e) 1.000, (f) 0.661 respectively. 88

4.6 Demonstrates some sample results where algorithm failed to yield an accurate centromere location. The detected centromere location (selected candidate) is depicted by a yellow dot while the segmented outline is drawn in blue. These results reported CBCC measures of (a) 0.368, (b) 0.066, (c) 0.655 respectively. 89

A.1 Comparison between Distance Potential (DT Based) model (top) and the GVF model (bottom). Each model depicts the initialization of the contour and convergence with each iteration (on left) followed by the final contour result after 100 iterations(on right) [5]. (Reproduced with permission from Prof. Jerry L Prince) 104

A.2 Comparison between Distance Potential (DT Based) vector field (left) and the GVF vector field (right) [5]. (Reproduced with permission from Prof. Jerry L Prince) 105

A.3 above: Steps of the DCE process of shape simplification, below: the same steps when initiated by adding boundary noise to the same image [6]. (Reproduced with permission from Prof. Longin Jan Latecki) 107

A.4 A representation of two line segments and the used angle measurements in the DCE process, which can be used to explain the rationale behind the used 'relevance measure'. 108

A.5 Shape variations of polygon partitions and the effects of turn angle and arc length to the relevance measure. 109

A.6 Two different separating hyperplanes derived for the same training data points. Figure A.6(a) contains the separating hyperplane with the optimal margin. The new data point given by the yellow square will be classified differently in these two cases. 113

A.7 The framework for deriving the geometric margin of a support vector machine classifier. 114

Acronyms and Symbols

1D, 2D	*1 Dimensional, 2 Dimensional*
A	*Adenine*
ADCI	*Automated Dicentric Chromosome Identifier*
ANN	*Artificial Neural Networks*
ANOVA	*Analysis of Variance*
C	*Cytosine*
CBCC	*Candidate Based Centromere Confidence*
CI	*Centromere Index*
DAPI	*4',6-diamidino-2-phenylindole*
DCE	*Discrete Curve Evolution*
DNA	*DeoxyriboNucleic Acid*
DT	*Distance Transform*
FISH	*Fluorescence In Situ Hybridization*
FITC	*Fluorescein Isothiocyanate*
G	*Guanine*
GF	*Goodness of Fit*
GVF	*Gradient Vector Flow*
KKT	*Karush Kuhn Tucker*
MAT	*Medial Axis Transform*
PC	*Parametric Curve*
PCA	*Principal Component Analysis*
RGB	*Red, Green and Blue*
ROI	*Region of Interest*
SLT	*Statistical Learning Theory*
SVM	*Support Vector Machines*
T	*Thymine*

Chapter 1 Introduction

The development of image processing techniques for analyzing human metaphase chromosomes can be seen as the key in speeding up many cytogenetical diagnosis processes while optimizing the use of the scarce resource, expert time.

A human chromosome is comprised of DeoxyriboNucleic Acid (DNA) along with protein. The DNA is primarily responsible for genetic inheritance and behavioral patterns of a human being. The genetic makeup and the familiar physical resemblance of a human chromosome is achieved due to the genetic condensation during cell division (mitosis). Therefore by studying the chromosome structure during mitosis, cytogeneticists can identify genetic disorders caused by genetic translocation, deletion, trisomy, monosomy and radiation exposure etc. Many chronic diseases are caused by these genetic deformations and can be diagnosed by analyzing chromosome cell images. Therefore the study of human chromosomes and their structure is of utmost importance in clinical diagnosis. Although cell preparatory techniques such as non radioactive Fluorescence In Situ Hybridization (FISH) have been used to assist this diagnosis process by providing the cytogeneticist with information regarding the present location of a known DNA sequence in a selected chromosome, the diagnosis process can still be tedious and time consuming [7]. A typical lymphocyte slide of a given patient on average can yield up to 500 cell images when imaged through a light microscope. Perusing through hundreds of cell images for one patient even at triage stage can be tiresome and can lead to operator fatigue. Therefore, manual analysis and diagnostic processes are tedious and tiresome for experts and also are limited by the the number of experts available.

With increasing use of digital microscopy for cytogenetical diagnosis, high resolution digital images are becoming readily available for the diagnosis process. With the adaptation of various staining methods and cell preparation technologies, the list of diseases that can be diagnosed also increases. One such technique is termed Fluorescence In Situ Hybridization (FISH), which places probes as markers for certain genetic sequences within the chromosome body. In Karyotype analysis, the expert needs to be presented with an annotated karyograms in order to diagnose effectively for chromosome abnormalities. Similarly in radiation dosimetry, the number of dicentrics (radiation dosage) needs to be counted accurately in order for the medical expert to prescribe a chain of remedies for the patient. This can grow into an even severe issue during the aftermath of a nuclear event where millions of patients would need these services from a handful of experts within a small window of time. Therefore with increasing capabilities of computer systems, developing a set of image analysis algorithms for analyzing chromosomes and aiding in diagnosing is a tempting task. This can certainly increase the efficiency of the diagnosis process while optimizing expert time usages. The research reported in this dissertation is part of a combined effort in developing a set of algorithms for detecting dicentric chromosomes for radiation biodosimetry.

Many methods have been tried over decades in order to fill this void for a set of algorithms to analyze human metaphase chromosomes. However, coming up with a set of algorithms to reliably detect salient features of human chromosomes remains a challenge to date due the morphological variations of the chromosome structure. The morphology and length of chromosomes within diploid human cells can vary between cytogenetic preparations depending upon the methodology used to grow and analyze the cells. The clinical reasons mostly governs the methodology or steps taken during cell preparation. If subtle structural chromosome abnormalities involving a single chromosome band are suspected as in certain inherited genetic disorders then methods (such as addition of DNA intercalating agents, reduced colcemid time, cell cycle synchronization, 3-4 day lymphocyte culture) that reduce chromosome condensation or arrest chromosomes in an earlier stage of metaphase are utilized (referred to as prometaphase or high resolution cytogenetic analysis). If numerical chromosome abnormalities or low frequency large structural acquired abnormalities such as dicentric chromosomes are suspected as in certain cancer specimens or biodosimetry samples then methods (such as prolonged colcemid time and concentration; 2 day cell culture) that increase the number of cells in metaphase at the cost of chromosome length are used. Aside from the obvious differences in chromosome length between the two types of methods, shorter more condensed chromosomes often have separated or distinct sister chromatids on each arm and fewer chromosomal bands are evident. Furthermore, the cell preparatory method and steps also depends on the type of laboratory which is utilized for the test. For an instance, cells prepared at a cytogenetic laboratory (for the diagnosis of many genetic abnormalities) would be in general lengthy compared to ones prepared at a biodosimetry laboratory (to calculate the radiation exposure dose of victims of a mass radiation event). Even environmental factors such as ambient temperature and humidity influences the shape variations in human metaphase chromosomes and a minute change in anyone of these factors can be represented in the shape of the chromosomes on the slides.

Despite these differences however, the primary constriction or centromere, which is the most constricted region of a chromosome, to which the spindle fiber is attached during mitosis (cell division) [8] remains evident on the chromosomes. Detection of the centromere involves in segmentation of the chromosome as well as identifying salient points such as the chromosome end points. Furthermore the accurate centromere localization can be used to directly identify the chromosome type and with additional information can lead to identifying the chromosome number in the cell as well. Therefore, accurate detection of the centromere location can be considered as a key element in a reliable chromosome analysis algorithm. However, detecting the centromere can be a challenging task even to the trained human eye. Irregular boundary conditions (especially in Giemsa stained chromosomes) as well as bent chromosomes can make the detection algorithm miss the constriction.

The ability to detect centromere locations can be extended into the field of radiation dosimetry that is required to detect dicentric chromosomes which are metaphase chromosomes with an additional centromere location as a bi product of radiation exposure. Since both centromere locations share similar characteristics, a set of features designed to detect the primary centromere in principle should be applicable to detect the secondary centromere as well. Once a set of suspected centromere locations are established, machine learning techniques can be used to reject dicentric false positives using features that captures the similarity between the two detected centromere locations and physical constraints such as the minimal distance between the two centromeres. Therefore the development of an algorithm for accurate centromere detection which is compatible with high morphological variations of chromosomes from multiple staining methods provides a good foundation for detecting dicentric chromosomes.

The majority of existing methods for chromosome analysis and centromere localization involve first calculating the chromosome centerline which is then used as a basis of measurement for calculating the thickness of the chromosome in order to detect the centromere location [9]. Generally the handful number of methods that do not utilize the centerline require some special cell preparation techniques for them to operate properly. Although centerline based methods do perform better than other methods, all existing methods manifest one or more of the following shortcomings.

- Suited to work with only a given type of staining / cell preparation method. The algorithms are designed to utilize some of the features specific to that particular staining / cell preparation method and therefore would not perform well for other types of analysis problems.
- The width constriction on higher banded chromosomes can be missed easily due to bends or noise on the chromosome boundary, while chromosomes with sister chromatid separation tends to mislead the width profile calculation near the telomeric region.
- The centromere detection result is not accompanied by any indication of either a measure of confidence or a measure of uncertainty. Cytogeneticists often makes critical decisions based on these detection results and false positives and negatives can have very high life threatening consequences.Therefore a measure of confidence in the localization would provide the expert with some insight into the detection process as opposed to a binary decision.

In this research, we explore image processing techniques and combinations to derive or compute various information from chromosome cell images. This information can be directly or indirectly used for clinical diagnosis and therefore can drastically reduce the time per patient. Since centerline based methods tend to perform better than other methods, we have proposed an algorithm which utilizes the centerline simply to divide the chromosome contour into two nearly symmetric partitions instead of using it as a basis for measurements. This approach prevents the possibility of boundary irregularities adversely affecting the centerline and therefore making the width profile measurements noisy. Once the contour is segmented, we then utilized a Laplacian based thickness measurement algorithm where intensity was integrated through a weighting scheme to bias the thickness measurement trace lines into homogenous intensity regions known as chromosome bands. The algorithm is capable of partitioning the telomere region, detecting evidence of premature sister chromatid separation and then correcting for the artifact. Finally, a classifier was trained where the distance from the separating hyperplane was then used as measurement of goodness of fit in order to find the best centromere candidate from the pool of candidates for a given chromosome. A Principal Component Analysis (PCA) was performed on the features that were created for centromere localization in order to gain some insight into the contribution from each feature to the overall variance of the feature set. A metric called 'Candidate Based Centromere Confidence' (CBCC) was introduced which represents the confidence in the selected centromere candidate. This provides the expert with useful information which he/she can then use in the diagnosis process. The proposed algorithm is designed to work with multiple staining methods and preparation procedures and is tested with a data set containing both Giemsa and DAPI stained cell images collected from multiple sources which were prepared for distinct clinical reasons. During preliminary testing, the proposed method was observed to be more accurate and statistically significant compared to a centerline based method [1]. We have tested the algorithm on a larger data set for further validation of its performance.

1.1 Contributions

The objective of this research is to develop a set of algorithms that can analyze human metaphase chromosomes originating from multiple sources with multiple staining methods. The algorithm is intended to work with chromosomes with high morphological variations and in the presence of premature sister chromatid separation [1]. The centromere detection is identified as a basis for measuring the performance of the proposed algorithm. Also a metric to yield the confidence in a given centromere detection was proposed. With the inclusion of the metric, the proposed algorithm provides experts with additional information regarding the detection process on top of the typical detection result.

The following are the main contributions of this research divided into 3 categories based on the functionality of the contribution.

[1] Please refer Chapter 2, page 23 - 27 in Concepts of Genetics by Klug and Cummings [10] for more details

Improvements to the thickness measurement of chromosomes

- The use of Laplacian algorithm for detecting the width of human chromosomes.
- Incorporating intensity into the Laplacian framework for guiding the width profile measurement process more accurately.

Providing additional information relating to a given centromere detection

- Creating a measure for measuring the confidence in a centromere candidate detection using the distance from the separating hyperplane as a measure of goodness of fit to a label.

Improvements to the applicability of the algorithm

- A chromosome centromere detection algorithm is proposed which is compatible with all chromosome classes and multiple staining methods.
- Combining the advantages of both centerline based methods and their counterpart to come up with a hybrid solution for obtaining the feature profiles.
- A candidate based centromere detection approach facilitates the inclusion of acrocentric and submetacentric chromosomes into the analysis.

Some of the publications related to this dissertation are listed below,

Patent publication

1. US Patent - Centromere detector and method for determining radiation exposure from chromosome abnormalities, United States 8,605,981. PCT No.: PCT/US2011/059257

Journal publications

1. Akila Subasinghe A. et al. In review - ”Centromere Detection of Human Metaphase Chromosome Images using a Candidate Based Method”. In Biomedical Engineering, IEEE Transactions on (TBME), 2014.

2. Peter K. Rogan, Yanxin Li, Asanka Wickramasinghe, Akila Subasinghe, Natasha Caminsky, Wahab Khan, Jagath Samarabandu, Joan H. Knoll, Ruth Wilkins, and Farrah Flegal."Automating dicentric chromosome detection from cytogenetic biodosimetry data",Journal of Radiation Protection Dosimetry, 2014.

3. Akila Subasinghe A. et al. "Intensity integrated Laplacian-based thickness measurement for detecting human metaphase chromosome centromere location". In Biomedical Engineering, IEEE Transactions on (TBME), volume 60, pages 2005 2013, July 2013.

4. W. A. Khan, R. A. Chisholm, S. M. Taddayon, A. Subasinghe, J. Samarabandu, L. J. Johnston, P. R. Norton, P. K. Rogan, J. H. M. Knoll. "Relating centromeric topography in fixed human chromosomes to a-satellite DNA and CENP-B distribution", Cytogenetics and Genome Research

Conference publications

1. Akila Subasinghe A. et al. "Intensity integrated Laplacian algorithm for human metaphase chromosome centromere detection". In Electrical Computer Engineering (CCECE), 2012 25th IEEE Canadian Conference on, May 2012.

2. Rajeev Ranjan, Akila Subasinghe Arachchige, Jagath Samarabandu, Peter K. Rogan and Joan Knoll. "Automatic Detection of Pale Path and Overlaps in Chromosome Images using Adaptive Search Technique and Re-thresholding", International Conference on Computer Vision Theory and Applications, 2012.

3. Yanxin Li, Asanka Wikramasinghe, Akila Subasinghe, Jagath Samarabandu, Joan Knoll, Ruth Wilkins, Farrah Flega, and Peter Rogan. "Towards Large Scale Automated Interpretation of Cytogenetic Biodosimetry Data", International Conference on Information and Automation for Sustainability, 2012.

4. Akila Subasinghe A, Jagath Samarabandu , Joan Knoll, Wahab Khan and Peter Rogan."Accurately extracting the centerline from human metaphase chromosomes using image processing". Canadian Student Conference on Biomedical Computing and Engineering (CSCBCE), 2012.

5. Akila Subasinghe A, Jagath Samarabandu , Joan Knoll and Peter Rogan. "Automated metaphase chromosome centromere refinement using fuzzy inference

systems". Canadian Student Conference on Biomedical Computing and Engineering (CSCBCE), 2012.

6. Akila Subasinghe A. et al. "An accurate image processing algorithm for detecting fish probe locations relative to chromosome landmarks on dapi stained metaphase chromosome images". In Seventh Canadian Conference on Computer and Robot Vision (CRV), May 2010.

7. Akila Subasinghe A. et al. "An image processing algorithm for accurate extraction of the centerline from human metaphase chromosomes". In International Conference on Image Processing (ICIP), September 2010.

1.2 Thesis organization

In this chapter we have discussed the problem domain addressed by the proposed algorithm including its contributions to the literature. Chapter 2 will provide some insight into the existing literature and provide a basic description of background methods as well as the anatomy of a chromosome. The proposed algorithm will be explained in chapter 3 while chapter 4 will provide the results of the proposed algorithm in order to gauge its performance. Chapter 5 will provide a summary of the conclusions drawn from this research and also provide some future work warranted by the proposed method.

Chapter 2 Background

The main objective of this chapter is to layout the background information to provide context for the proposed algorithm described in chapter 3. This includes a brief introduction to human chromosome anatomy followed by a review of the existing literature and a theoretical background of some of the techniques required to comprehend the proposed method.

2.1 Introduction to human chromosomes

In every living organism (except some viruses), nucleic acid DNA (deoxyribonucleic acid) makes up the genetic material. DNA is essentially a double stranded molecule organized as a double helix which stores the hereditary units known as genes (see figure 2.1). The smallest unit in this double helix is known a nucleotide which is composed of Deoxyribose (a 5- Carbon sugar molecule), a phosphate and one of four nitrogen bases Adenine (A), Cytosine (C), Thymine (T) and Guanine (G). Deoxyribose and phosphate bond together to create a twin backbone in the reverse order on either side of the helix, while connection between the two strands are made by the relatively weak nitrogenous bonds. These bonds happen in a very specific order wherein, Adenine (A) only connects with Thymine (T) and Cytosine (C) only connects with Guanine (G) and vice versa. Each of these connections make up a single base pair. On average, a human chromosome contains about 100 million base pairs of DNA [11]. The chromosome can contain non-genic regions on top of the vast abundance of genes. During mitosis (and meiosis), the diffuse network of genetic material in the nucleus known as chromatin condenses and folds up while giving the chromosome its characteristic shape temporarily only to return to the original state towards the end of mitosis. During mitosis, chromosomes must ensure that the DNA matter is separated equally to daughter nuclei during mitosis while maintaining the integrity of the genome [10].

When imaged using a light microscope during some stages of mitosis, a healthy human cell image should contain 46 chromosomes. This consists of 22 pairs (autosomes) of chromosomes (a number referred as the haploid number in a cell) and two sex chromosomes X and Y. The presence of two XX chromosomes normally represents a female while XY would normally specify a male. The 44 autosomes are numbered from 1-22 in descending order of the length, size and the centromere position of each of these pairs [12]. Each chromosome in a pair with nearly identical length and centromere placement are called 'homologous chromosomes'. The karyotype of a G-banded image given by figure 2.2 shows this nearly identical length and centromere placement except for the two sex chromosomes (in male subjects). Furthermore, homologous chromosomes carry the same gene sites along their lengths and therefore have similar genetic potential. In humans (as with any sexually reproducing species), one of these homologous chromosomes is derived from the paternal parent while the other from the maternal parent.

A human metaphase chromosome has the following functional regions (depending on the chromosome type) which can be visually identified as,

- The *centromere*
- The *telomere*
- The nucleolar organizer regions

During this research we were mostly interested in the centromere and the telomere regions.

2.1.1 The centromere

Centromere is the condensed or constricted location which holds the two sister chromatids together in place. It acts as the site where the spindle fibers attach to during mitosis [8]. This is the location that is critical in chromosome segregation and cell division in both meiosis and mitosis. A mistake in the meiosis stage can yield incorrect number of chromosomes in cells and can lead to disorders such as the 'Down syndrome'. In many groups of chromosomes, this region in general can be observed as a clear constriction in relation to the width profile of the chromosome. The centromere positioning in these chromosomes also determines their shape during a later stage of

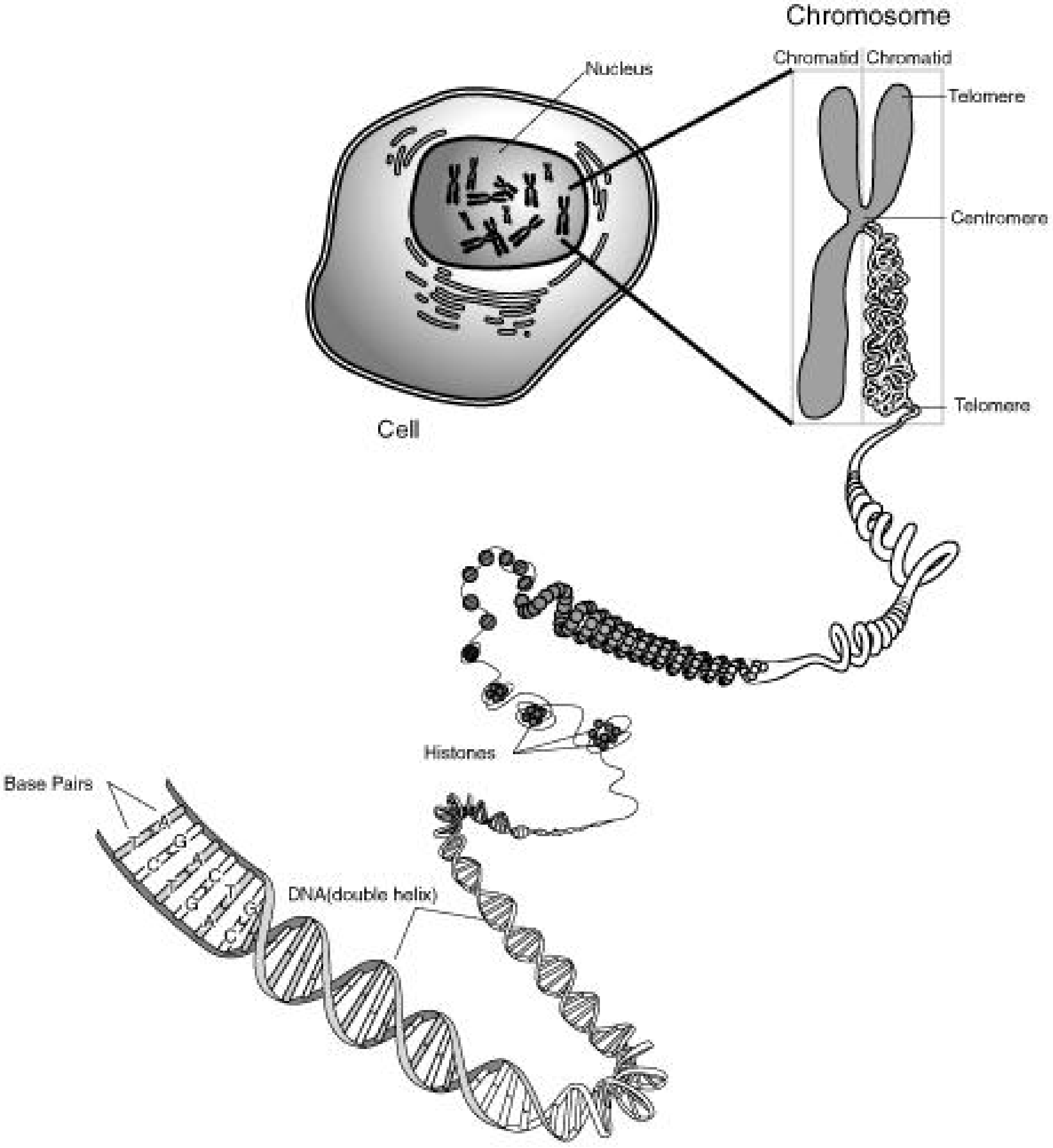


Figure 2.1: An illustration of the structure of a chromosome in context of the cell (Source- http://commons.wikimedia.org/wiki/File:Chromosome.gif).

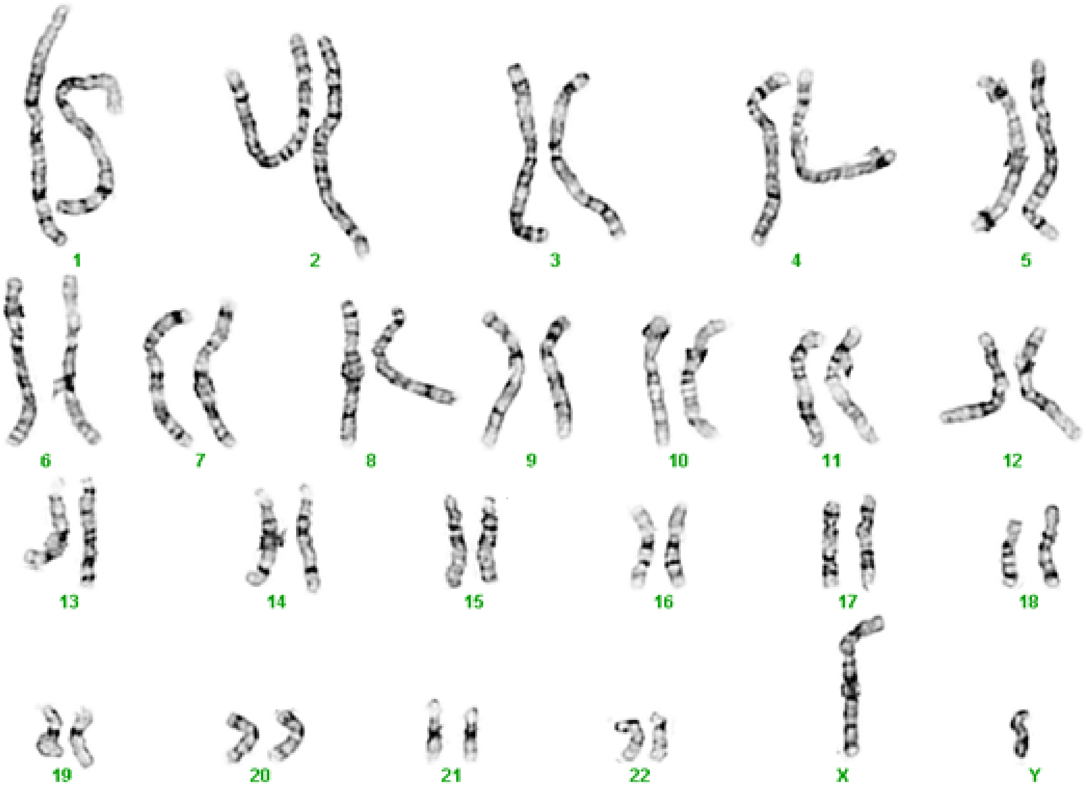
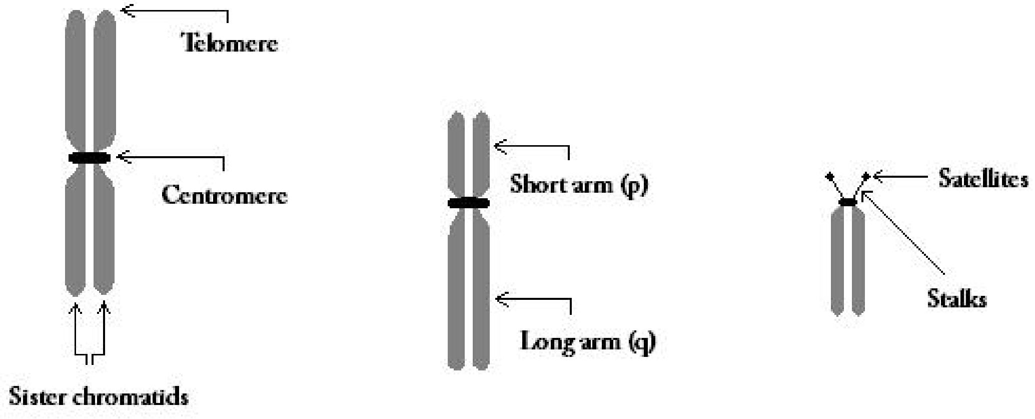


Figure 2.2: A karyotype analysis end result with all 46 chromosomes organized according to their chromosome groups and types (Reproduced with permission from Dr. Joan Knoll and Dr. Peter Rogan).

mitosis called the anaphase. Furthermore, human chromosomes can be grouped into 3 categories based on the location of the centromere with respective to its ends as follows,

- *Metacentric*
- *Sub-metacentric*
- *Acrocentric*

The centromere of metacentric chromosomes are located near the middle of the chromosome while in acrocentric chromosomes, it is near one of the end points. Sub-metacentric chromosomes have the centromere between the middle and one of the end points of the chromosome. All three different types of chromosomes, including the acrocentric type (with 'nucleolar organizer regions' or 'satellite stalks') are depicted in figure 2.3. Additionally chromosomes that do not possess a functional centromere are called 'acentric', while those with two centromere locations are called 'dicentric'. Most of the time only one of the two centromere locations in a dicentric chromosome is active during mitosis.



(a) Metacentric (b) Submetacentric (c) Acrocentric

Figure 2.3: The structural components of metaphase chromosomes of all three categories : the metacentric, sub-metacentric and acrocentric

2.1.2 The telomere

The name telomere is derived from the Greek term 'telos' which means the 'end'. Apart from the centromere (see figure 2.3), the telomere can be regarded as the second most important structure of the chromosome. Located at the ends of the chromosomes, its primary function is to prevent the chromosome from interacting with other chromosomes in the cell by rendering the ends of the chromosome inert [12]. Therefore telomere regions do not fuse with one another or with other broken ends. This is important since the ends of broken DNA molecules tend to fuse together easily. Preventing such unwanted fusions is critical in cell propagation in the organism [13].

2.1.3 The centromere index (CI)

The *centromere index* (CI) is a measure based on the location of the centromere (see section 2.1.1) with respect to the ends of a chromosome. The value of this index is defined using figure 2.4.

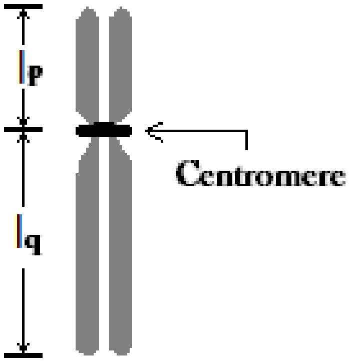


Figure 2.4: The lengths used for calculating the centromere index of a given chromosome.

Let l_p and l_q respectively be the lengths of the short-arm (p-arm) and the long-arm (q-arm) of the chromosome. Then CI is the ratio between the short-arm length to the total length of the chromosome, and is stated as,

$$CI = \frac{l_p}{l_p + l_q} \tag{2.1}$$

Therefore, it can further be observed that the CI value lies in the range of 0 and 0.5. For each chromosome in a cell (based on the chromosome number), the CI value must fall within a certain small interval. Therefore, the CI value is an important value that can be used to assist chromosome identification and classification. Table 2.1 below provides CI value ranges for all chromosomes in a human metaphase cell [2].

Table 2.1: Typical range of centromere index (CI) for each human chromosome [2]

Chromosome Number	**CI value**	**Chromosome Number**	**CI value**
1	0.45 - 0.50	13	0.13 - 0.22
2	0.35 - 0.42	14	0.13 - 0.22
3	0.44 - 0.50	15	0.13 - 0.22
4	0.24 - 0.30	16	0.41 - 0.45
5	0.24 - 0.30	17	0.28 - 0.37
6	0.34 - 0.42	18	0.23 - 0.33
7	0.34 - 0.42	19	0.42 - 0.50
8	0.33 - 0.38	20	0.41 - 0.50
9	0.32 - 0.40	21	0.22 - 0.30
10	0.30 - 0.37	22	0.22 - 0.30
11	0.35 - 0.45	X	0.36 - 0.41
12	0.24 - 0.30	Y	0.28 - 0.34

2.2 Review of existing algorithms

With the use of cytogenetical analysis methods, the demand is rising for automated microscopy systems that can increase throughput while not compromising the accuracy. This is specially the case since the speed of the diagnosis process is highly dependant on the time required by a trained cytogeneticist to examine the chromosome images. Therefore, a significant amount of research has been carried out to automate these processes in order to present the data in a better way to the experts so that they speed up the diagnosis process. One such attempt where many research publications have being carried out is Karyotyping, where the algorithm provides an annotated list of all the 46 chromosome in a Giemsa - banded human cell image. Fluorescence in situ-Hybridization (FISH) is another such attempt of analysis where the algorithm is required to detect a fluorescent probe hybridized in the DAPI stained chromosome body and provide the expert information regarding its positioning with respect to the chromosome structure. In radiation dosimetry, the objective of the algorithm is to detect and count the number of dicentric chromosomes in the cell image, which are by-products of radiation exposure. An expert can utilize this information in order to diagnose the amount of radiation exposure and then to prescribe a suitable remedy accordingly. All these methods and approaches rely on obtaining the following information,

- An accurate segmentation of the chromosome
- An accurate localization of the centromere location
- A mechanism to detect supplementary information such as banding patterns, length of the chromosome etc.

In this section we have provided a detailed literature review of the existing methods of segmentation and centromere detection.

2.2.1 Segmentation methods

Image segmentation can be defined as the process that partitions a given digital image into many non-overlapping (disjoint) regions which correspond to individual objects [14]. This is an essential step in isolating the chromosome from the cell image.

Some of the approaches relied on manual segmentation of the chromosomes where a technician had to mark the outline of the chromosome by hand [15], [9]. This was a very time consuming process which produced subjective results for the segmentation of each chromosome.

Chromosome cell images generally present a reasonable degree of contrast in terms of intensity between the objects and the background. In other words, the image histogram is bi modal and separable. Therefore many of the existing approaches tend to employ intensity based thresholding in order to segment images [16]. In one approach Sadina and Mehmet simply segmented the chromosomes using a fixed threshold value set at 0.9 of the normalized intensity of the Giemsa banded image [17]. Since G-banded image background intensity is higher than the intensity values within the chromosomes, they marked intensity values less than 0.9 as the object and the rest as the background. Since general intensity characteristics can change from one cell image to another, having a fixed intensity level would yield highly inconsistent segmentation results.

Many of the automated segmentation techniques are performed based on global thresholding for different staining methods, where the algorithm calculates a suitable intensity value in order to separate the bimodal histogram into two segments. Popescu et al. performed global thresholding on G-banded images using an algorithm termed 'Otsu's method' [18] and used this segmentation result as the initial stage of segmentation in his approach [19]. Similarly, Wolf et al. utilized the same algorithm for segmentation on DAPI stained images used for FISH [20]. In another approach a thresholding algorithm called 'Ketler's method' [21] was used to globally threshold the chromosome cell image [22]. Some authors resorted to operate directly on the image histogram for obtaining the segmentation result. In one such approach, Ji [23] segmented chromosomes by applying a threshold value based on the smoothed histogram of the chromosome image. This initial value was selected to be the value where the intensity gradient (slope) of the histogram becomes zero. Then he re-thresholded the first result with a higher threshold value.

Since thresholding is a point operation where each pixel is labeled based on its intensity value, this method is prone to creating noise in the segmented binary image. Therefore many authors have attempted to employ both pre processing and post processing methods to minimize this phenomenon. Gajendran and Rodriguez proposed the use of 'hysteresis thinning' (used in the 'Canny edge operator') as a post processing step on the threshold output in order to reduce some of this noise content [24], [25]. Wang et al. utilized pre processing approach where a median filter was used to remove some of the noise in the original image which could lead to a noisy segmentation result [26], [27]. Then, the thresholded image was subjected to 4-connected component labeling to remove isolated noise in the binary image.

Despite the fact that some of these approaches were successful in significantly removing noise in the binary segmented output, the global thresholding approach in general remains highly dependent on the lighting conditions in the image. Uneven illumination in the cell image can cause the thresholded objects to be noisy and even discontinuous at some locations. Furthermore, chromosomes could cluster together as one blob if the threshold value is set incorrectly or on the other extreme, even could break the chromosomes into multiple segments. In some of the approaches, the threshold value was set locally (adaptively) based on the immediate vicinity of the chromosome to possibly solve this problem. Enrico et al. attempted this by dividing the cell image into many tessellations with manually set fixed sizes and applied thresholding on each of those regions of interests (ROIs) [28].

In general, segmentation using thresholding is highly sensitive to both quantization errors as well as to intensity fading around chromosome boundary regions. This tends to create a noisy object boundary on the binary output and therefore fails to represent the intricate shape variations of the corresponding chromosome in the cell image. However, local thresholding could be considered as a very effective segmentation step when followed by a refining step.

Few methods can be found in literature where parametric active contour models such as Gradient Vector Flow (GVF) have been utilized for segmentation. GVF is an improvement of the standard active contour model [29], where limitations such as the small capture range and lack of convergence into boundary concavities were addressed [30]. The works of Britto & Ravindran and also Li et al. has reported significant improvements in chromosome segmentation by using the GVF snake model [31], [32]. However, GVF being a parametric active contour, the global minima is not guaranteed unless the control points are initialized in the vicinity of the desired contour (even with the improved capture range). Therefore the contour could converge to an unwanted local minimum such as a chromosomal band (which has a strong intensity gradient) or even to the contour of another chromosome.

2.2.2 Centromere detection methods

Accurate detection of the centromere location in a chromosome is a critical step in any automated diagnosis process. The location is characterized by a constriction in the width and sometimes by a relatively lower intensity region within the body of the chromosome. The region of darker intensity depends on the specific staining method and the approach taken by the technicians when preparing the samples. Therefore, the width constriction can be considered as the more universally evident feature of the centromere location. Centromere localization methods in the literature vary mainly based on the methodology of obtaining the width profile of the chromosome. Therefore, methods for centromere detection can be divided into the following two categories,

- Methods that first calculates the centerline of the chromosome : The morphological centerline of a closed object are defined as the set of all points which are centers of circles (in 2D case) that are tangent to the shape at more than one point and that contain no other tangent circles [33]. Once the centerline is calculated, it will be used as the basis for calculating the width and/or intensity profiles of the chromosome. This is often performed by creating a trellis structure along the longitudinal axis of symmetry.

- The few methods that do not rely on a calculated centerline as the basis for measurement of the width of the chromosome : The applicability of these methods are often restricted by the necessity for special preparation techniques or by morphological conditions.

Medial Axis Transform (MAT) and morphological thinning are the most commonly attempted methods of finding the centerline in the literature. Medial axis transform or skeletonization attempt to reduce the segmented object into a set of pixels which preserves the extent and connectivity of the original object. One such attempt was made by Wolf et al., in which the binary segmented image was subjected to morphological closing (dilation operator followed by the erosion operator) before applying Medial Axis Transform (MAT) to find the centerline [20]. The rationale behind applying the closing operator was to smoothen the object boundary before skeletonization. The author resorted to manual user interaction based corrections when any spurious branches were present. Therefore, this process is far from being autonomous. Moradi & Saterahdan proposed a better approach in which the problem of having bifurcations (in the skeleton) towards the ends of the chromosome was solved [9]. They took the median line of the triangle formed by the two skeletal segments and the chromosome boundary at the telomere regions. Yet, this method also fails if the skeleton gives spurious branches away from the telomere regions. In another attempt, Stanley et al. calculated the feature profiles using a trellis structure based on a centerline derived using MAT [34]. In all these methods, the main weakness is the accuracy of the centerline, which can be quite unreliable due to the occurrence of spurious branches. Furthermore, MAT provides a set of points in space, rather than a parametric curve that could effectively and easily be used for further calculations.

Thinning on the other hand creates less spurious branches compared to skeletonization. These methods are often accompanied by a method for end point extension since they remove data points from the extreme ends of the centerline [4], [35]. In one such attempt, Wang et al. applied morphological thinning to the segmented binary object and then sampled it with a 5-pixel interval. Then these points were interpolated to obtain the chromosome centerline [36],[27]. Gajendran & Rodriguez applied median filtering to the digital cell image prior to obtaining the thinned centerline of the chromosomes [24]. Some of the approaches have utilized iterative thinning algorithms which preserve the ends of the centerline unlike previous thinning methods. In one such approach Somasundaram and Kumar used a method called the 'stentiford thinning algorithm' in order to obtain the complete centerline of the chromosome [22]. Yet, irregular boundary conditions which are commonly observed in Giemsa stained images can introduce spurious branches in the thinned centerline.

We have previously proposed an algorithm to calculate the centerline with no spurious branches irrespective of boundary irregularities or the morphology of the chromosome [1]. Mohammad proposed an approach where he used our aforementioned algorithm to derive the centerline and then used a curvature measure to localize the centromere location instead of the width measurements [37]. Despite the lack of spurious branches, irregular boundary conditions (boundary noise) adversely affect the centerline derived through these methods. Measurements performed on a noisy centerline can easily lead to false centromere localization.

Several attempts have been made in order to find suitable methods without using skeletonization or thinning due to their aforementioned inherent weaknesses. Jim

Piper and Erick Granum [38] proposed a two stage approach to find the centerline in which they first determined the orientation of the chromosome by calculating the minimum width enclosing rectangle. Then, if the chromosome is not highly bent, it was rotated such that the orientation was vertical and mid points of the horizontal chromosome slices were connected together to obtain the centerline which was then smoothed to get the 'poor man's skeleton'. But, if the chromosome is bent, they performed a conventional skeletonization algorithm. Yet, the problem with this approach is the spurious branches that occurred with the conventional skeletonization process. In another approach [39], chromosomes were sampled into scan lines of different inclinations and after selecting proper cross-sections, the selected mid points were combined to obtain an approximate centerline. The drawback of this method is that it attempted at getting a polygonal approximation of the centerline instead of the centerline itself. Results were poor when the segmented chromosome boundaries were irregular in shape, which is a common occurrence in medial imaging. Gunter Ritter [40] proposed a method which was based on finding the dominant points of the chromosome. But, results were not reliable when it was applied to highly bent and blurred chromosomes.

Due to the non rigid structure of bent chromosomes it has been one of the most challenging aspects in developing an algorithm for centromere localization. The bend points can introduce spurious branches to morphological operators such as thinning or MAT and also cause the centerline based method's trellis structures to miss the actual centromere location. Most chromosomes bend at the centromere location and therefore exacerbate these false localizations. Some of our previous work have focused on getting a centerline without spurious branches [41] while retaining the original shape and orientation of the chromosome while Piper et al. rotated the chromosome to align the centerline vertically [38]. Another such centromere detection approach attempts to straighten bent chromosomes prior to detecting the width minima [42]. The straightening process analyzes the vertical and horizontal projection vectors of the chromosomes calculated at a set of rotation values in order to find the best rotation to align the centerline vertically [43]. However this algorithm works only with chromosomes limited to one bending center and is expected work well only on chromosomes in group E (chromosome numbers 16-18).

There has been some research work in the literature, where the centromere detection does not involve in finding the centerline of the chromosome [36]. Mousavi [44] assigned a membership value for each pixel of DAPI (4',6-diamidino-2-phenylindole) and FITC (Fluorescein Isothiocyanate) images (with centromere probes) based on an iterative fuzzy algorithm. However, this method required the use of centromere probes to mark the location of the centromere. Another work carried out by Moradi [15] and similarly by Faria [45] (on chromosomes of fish) attempted to find the centromere location by getting the horizontal and vertical projection vectors of the binary segmented chromosomes. Both methods did not perform well on acrocentric chromosomes as well as on chromosomes with a bend greater than 90^0 degrees.

Considering the above limitations and shortcomings, there exists the need for a centromere localization algorithm which can perform well with any morphological variation as well as with multiple staining methods. Furthermore, none of the approaches in literature can correct for artifacts such as sister chromatid separation. Similarly, we are yet to encounter a centromere localization algorithm which provides relevant supplementary confidence measurement values for each centromere localization.

2.3 Background methods

This dissertation employs a number of existing image processing and machine learning algorithms. A brief description is given below. For a more detailed description, the reader is referred to appendix A.

Gradient Vector Flow (GVF) snakes - This is a commonly used active contour model based segmentation algorithm. GVF uses an edge based static vector field as the external energy for evolving a set of points which constitutes a closed/open snake. This segmentation algorithm has a higher capture range and the ability to converge into boundary concavities better than the standard active contour models. Therefore, this algorithm was used to obtain smooth object boundaries of human metaphase chromosomes. A detailed description of the GVF snake algorithm along with a comparison with the distance based snake model is given by appendix A1.

Discrete Curve Evolution (DCE) - DCE is a polygonal shape simplification algorithm which evolves by iteratively deleting vertices of a given polygon based on a relevance measure. This measure captures the contribution of each individual vertex to the overall shape of the polygon. In this research, this algorithm was utilized to locate chromosome salient points in order to partition the object boundary. Appendix A2 provides a detailed description of the algorithm and the relevance measure along with the advantages and disadvantages of this approach for detecting salient points.

Support Vector Machine (SVM) - SVM is a powerful kernel based supervised learning technique. SVM maximizes the margin between the two classes using the training data set. This provides good generalization and therefore is more likely to perform well with unseen data. Furthermore, the use of kernels to map data into a higher dimensional space increases the probability of obtaining a better separation between the class labels. In this research SVM was used as a classifier in multiple learning problems including contour partitioning, shape analysis and centromere detection. In some instances, the distance from the separating hyperplane (geometric margin) was used as a measure of goodness of fit of a given sample. The basic framework of SVM along with the derivation of the classification problem is given in appendix A3.

Chapter 3 Proposed algorithm

Detecting abnormalities in the human metaphase chromosome structure is a key stage in the cytogenetic diagnosis process. Digital image analysis algorithms can speed up this process to effectively utilize valuable and scarce expert time. However, the existing algorithms in the literature can only operate on a limited range of shape variations that a chromosome can exhibit with a specific staining method. Therefore, an algorithm is proposed in this research which could operate with multiple staining methods and chromosome morphologies. The proposed algorithm is able to perform segmentation, extract the centerline, detect the centromere location and to detect and correct for sister chromatid separation. The algorithm also provides cytogenetic experts with a measure of confidence in a given centromere detection. It is developed and tested with both DAPI and Giemsa stained images and is readily adoptable to work with other staining methods.

The algorithm requires the user to manually pick a point within (or close to) each chromosome in order to proceed with the rest of the process autonomously. The algorithm assumes that the marked chromosome does not either touch or overlap with other chromosomes in the cell image. This assumption is reasonable due to the use of a content based ranking algorithm proposed by Kobayashi et al. in this approach [46]. The output of this algorithm was a ranked set of metaphase images where chromosome images that were spread well with minimal overlaps and were complete (contain all 46 chromosomes) were ranked higher. Typically from a given set of cell images, only the highest ranked 5% were selected for further processing. This is a critical step required to improve the accuracy of the proposed algorithm.

The proposed algorithm which is designed as a sequential set of processes, is depicted in the flow diagram given by figure 3.1. The user selected chromosome is first segmented out from the cell image. Next, the centerline of the chromosome is derived using the binary segmentation result. The algorithm next partitions the telomere regions of the chromosome in order to detect evidence of sister chromatid separation. If the presence of sister chromatid separation is detected, the proposed method corrects for the artifact. The correction is performed in order to obtain an approximately symmetric partitioning of the contour which is a prerequisite for the IIL (Intensity Integrated Laplacian) thickness measurement algorithm. The Laplacian based thickness measurement algorithm was improved by integrating intensity information to utilize chromosome intensity bands. Once the thickness measurements are calculated, the proposed method creates multiple candidates for the centromere location based on local minima. Next, the candidates are ranked and the best candidate is selected as the centromere location. The proposed method then calculates a measure termed 'Candidate Based Centromere Confidence' (CBCC) which yields the confidence of the centromere detection based on the candidates.

The proposed algorithm will be explained in the following five functional stages,

- Preprocessing and segmentation (discussed in section 3.2)
- Finding the chromosome centerline (discussed in section 3.3)
- Contour partitioning & correcting for sister chromatid separation (discussed in section 3.4)
- Laplacian based thickness measurement (discussed in section 3.5)
- Candidate based centromere detection (discussed in section 3.6)

3.1 The data set

The research was carried out as a part of a combined effort for developing a set of algorithms to perform dicentric chromosome detection. Samples of peripheral blood lymphocytes were prepared to obtain metaphase cells, then metaphase cells were stained with either Giemsa or DAPI, imaged and analyzed in laboratories at Health Canada (Dr. Ruth Wilkins), Atomic Energy of Canada Ltd (Ms. Farrah Flegal) and the University of Western Ontario (Dr. Joan Knoll, Pathology Dept). Figure 3.2 provides an example for the two staining methods used for this research. The complete data set used for developing and testing the algorithm discussed in this dissertation consists of 40 metaphase cell images including 38 from biodosimetry samples and 2 from clinical cytogenetic samples. The chromosome data set comprised images of 18 Giemsa stained cells and 22 DAPI stained cells. These metaphase images were manually selected from a pool of 1068 cell images. The main criteria of the selection

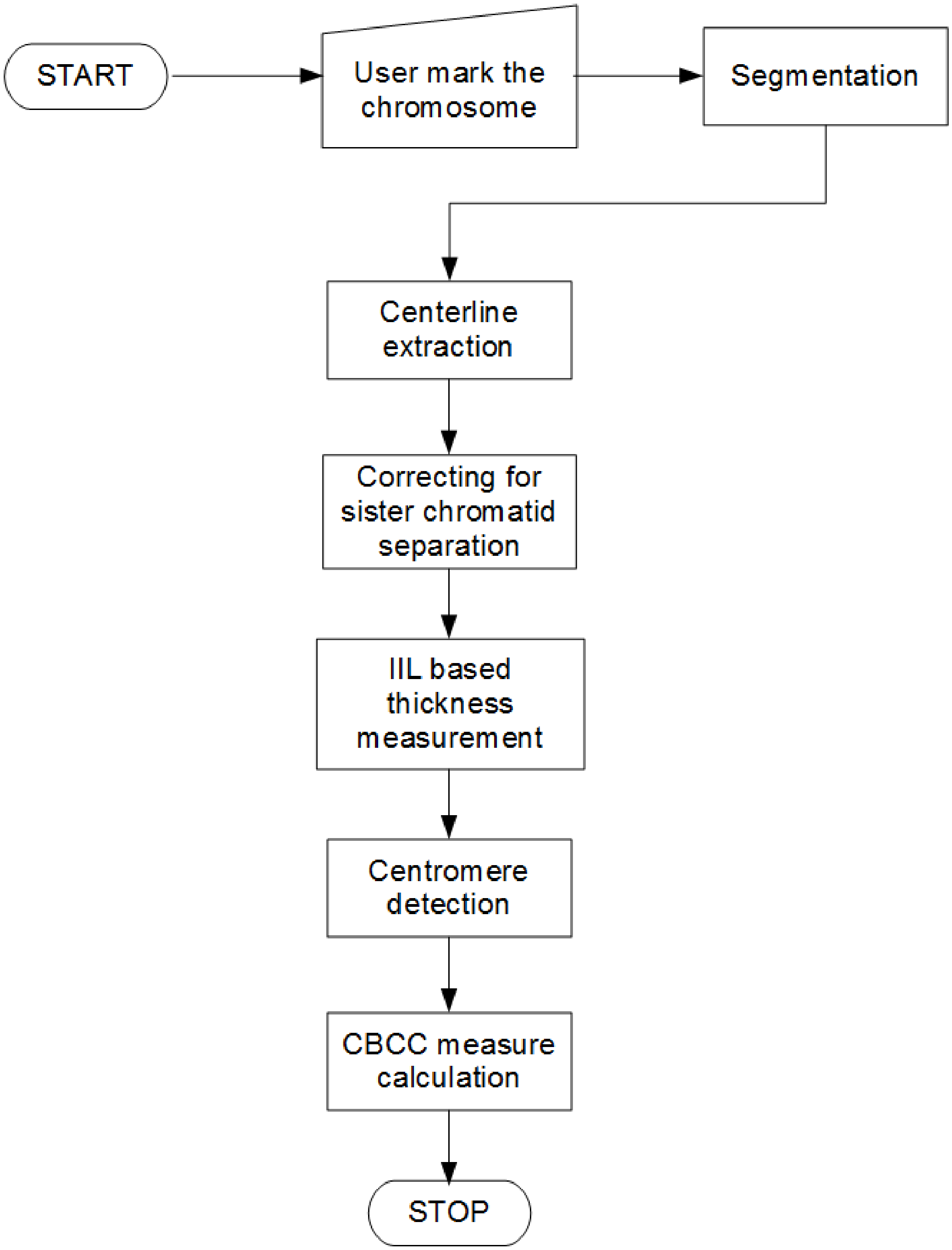
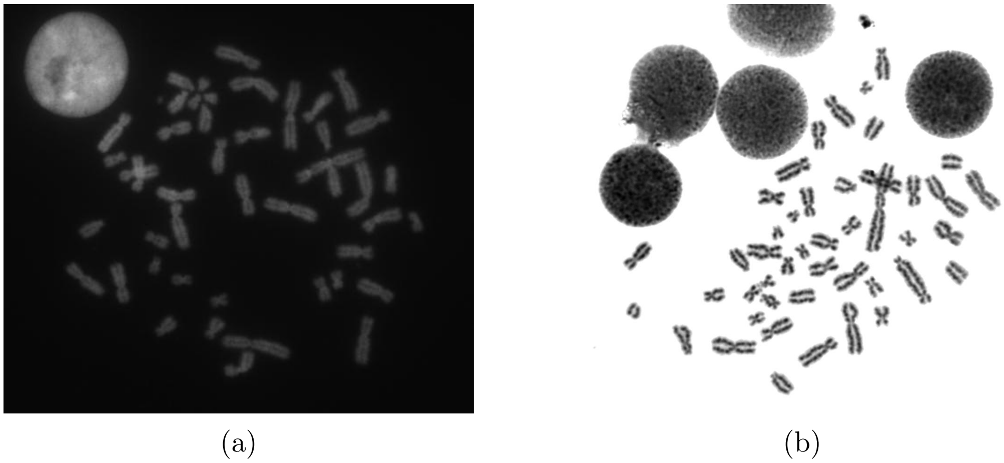


Figure 3.1: The flow diagram of the proposed method.

was to gather a representative sample of cells from both DAPI and Giemsa stained images with no bias to the length of the chromosomes. In the case of DAPI images, the selection was performed to include chromosomes with and without premature sister chromatid separation which captures a large degree of morphological variations. Furthermore, the selected cell images had a good spread of chromosomes (containing all 46 chromosomes) with minimal touches and overlaps which is a feature that enables the algorithm to extract more chromosomes from each cell image.


(a)

(b)

Figure 3.2: Shows two cell images with different staining methods. Figure 3.2(a) contains a cell image with DAPI staining while figure 3.2(b) is a Giemsa stained image.

The data collection process for the experiment was performed using the Matlab version of the algorithm while a converted and parallelized C++ version of the algorithm (termed ADCI - Automated Dicentric Chromosome Identifier) was developed and tested by the combined efforts of Mr. Asanka Wickramasinghe and Mr. Yanxin Li. For the data collection, the chromosomes were manually selected in order to pick all possible chromosomes with no touches or overlaps with neighboring chromosomes (judged visually). The interface required the operator to select a point within or in the vicinity of the chromosome of interest while the rest of the process was fully automated. Some control was given to the operator to hard segment the chromosomes in cases where separation was possible with minimal change of the threshold value (in cases of chromosomes that are barely touching each other). This was obtained using a thresholding factor (default value of 1.00) which was used to multiply the Otsu's threshold value (see section 3.2). The author was able to extract 1400 chromosomes in total from these images with no touching and overlapping present, which averages to 35 chromosomes per cell image. However, the thresholding factor was required to be changed only in 44 chromosomes out of the 1400 cases to the value of 1.05 in order to perform the separation. All the images were converted to gray scale (0 - 255) from the RGB (Red, Green and Blue) format for processing. The ground truth collection for training and testing all machine learning problems discussed in the dissertation were performed by the author.

3.2 Pre processing and segmentation

Chromosome metaphase images are often subjected to uneven illumination and could contain nuclei which appear as bright blobs under the microscope. Since these artifacts can adversely affect the segmentation process, each chromosome was processed individually. A user was required to mark a pixel within or close by the chromosome, which in turn was used to extract a fixed window containing the chromosome as the 'Region Of Interest' (ROI) for further processing. This window had to completely include the chromosome of interest while also including some portion of the background as well. The dimensions of this ROI was set to 201x201 empirically. This value was observed to be sufficient to include all chromosomes in the given data set (collected using the standard 100X magnification). However, if needed, the value may be changed to accommodate more elongated chromosomes in the future.

Chromosome metaphase cell images tend to contain pixels with limited range of intensity values (out of the possible 256 levels in the digital image). The effects introduced by the fluorescence light source also contributes to this feature. For example, a window extracted from the middle of a fluorescence microscopy image would on average have brighter (higher) intensity values compared to a window extracted from the corners of the image. Also, in terms of segmentation, a well spread bi-modal histogram intuitively would lead to better results. Therefore the intensities of this extracted ROI was normalized. The normalization was performed using a technique called the 'window center adjustment' (see figure 3.4) which calculates the highest and lowest intensity values and then scale them linearly in order to fit the possible range of intensity values of 0 - 255 by setting the window size to match the original intensity range. The increase of contrast on a DAPI stained chromosome window due to intensity normalization is depicted by figure 3.5.

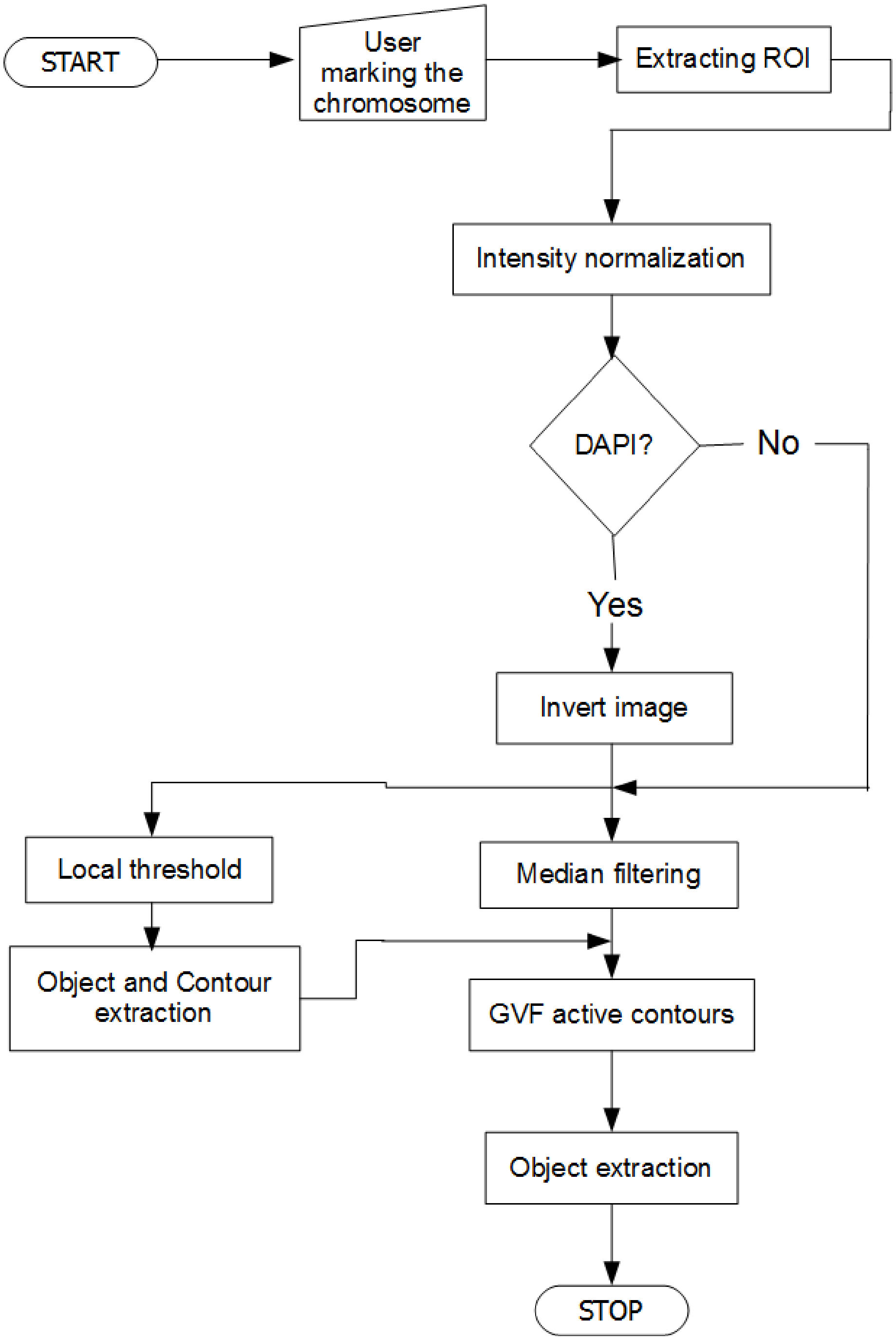


Figure 3.3: The flow diagram of the preprocessing and segmentation stage of the proposed method.

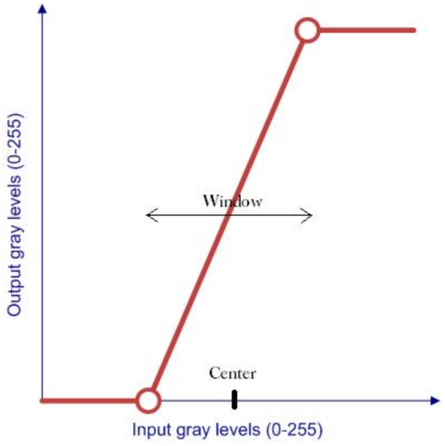


Figure 3.4: The window-center intensity mapping scheme which was used to map a certain intensity range (defined by the window and the center) to the full range (intensity levels 0 - 255).

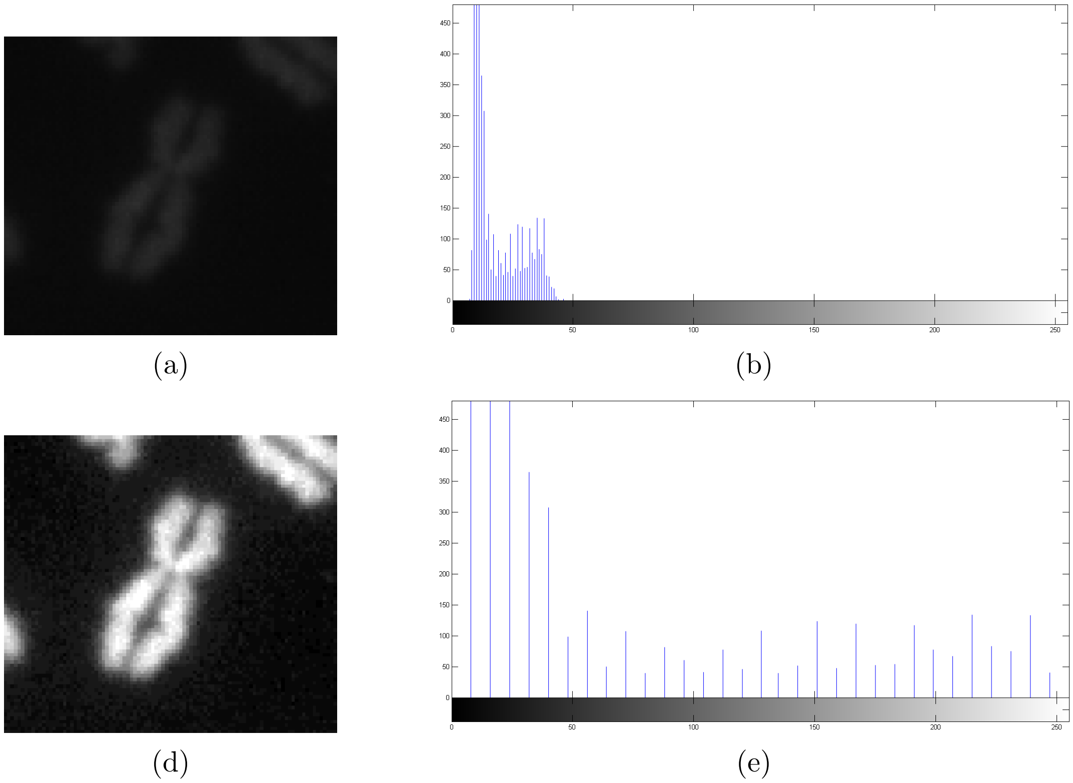
DAPI images are composed of chromosomes with brighter (higher) intensities with a dark (low) intensity background as opposed to Giemsa images which have darker intensity chromosomes in the bright background. In order to process images with both staining methods, intensity values of the DAPI cell images were inverted to obtain an appearance consistent with Giemsa stained images.

The initial stage of segmentation consists of thresholding the image using Otsu's method [18]. Otsu's method is a clustering algorithm that attempts to find the optimum threshold value that minimizes intra-class variance (background/foreground). The intra-class variance σ_c is calculated according to equation 3.1, where $q_{c1}(t)$ & $q_{c2}(t)$ are estimated class probabilities while $\sigma^2_{c1}(t)$ & $\sigma^2_{c2}(t)$ are their respective individual class variances. Once the threshold intensity value (T_{int}) is found, the image intensities are converted to a binary image (I_{bin}) using equation 3.2.

$$\sigma^2_c(t) = q_{c1}(t) * \sigma^2_{c1}(t) + q_{c2}(t) * \sigma^2_{c2}(t) \tag{3.1}$$

$$I_{bin} = \begin{cases} I_{bin} = 1, \text{ if } I_{x,y} < T_{int} \\ I_{bin} = 0, \text{ if } I_{x,y} \geq T_{int} \end{cases} \tag{3.2}$$

In binary thresholding, it is a common mistake to segment some neighboring chromosomes into a single binary object despite the presence of a visual separation.



(a) (b)

(d) (e)

Figure 3.5: Depicts an example of the effect of intensity normalization on a DAPI stained image window. Figure 3.5(a) & (b) depicts the original image window and the corresponding histogram. Similarly the figures 3.5(c) & (d) depicts the intensity normalized image window and the corresponding histogram.

This phenomena is due to the washed out intensity patches (called 'pale paths' in literature [19], [23]) evident between nearby chromosomes in metaphase cell image. The user was provided with means of separating such touching chromosomes through the use of thresholding factor (default value of 1.00). The threshold factor was used to multiply the threshold value calculated above using Otsu's method to perform a hard segmentation. Although this allowed gathering of more chromosomes from a given cell image, it is important to notice that excessive hard thresholding could have adverse effects on the object contour. In extreme cases, this can even break the chromosome into few segments. In this experiment, the thresholding factor was only increased to 1.05(a small increment) in approximately 3% of the chromosomes collected (44 chromosomes out of the 1400 collected). The use of the thresholding factor was a interim solution and is planned to be removed following the implementation of an algorithm for accurately separating touching and overlapping chromosomes.

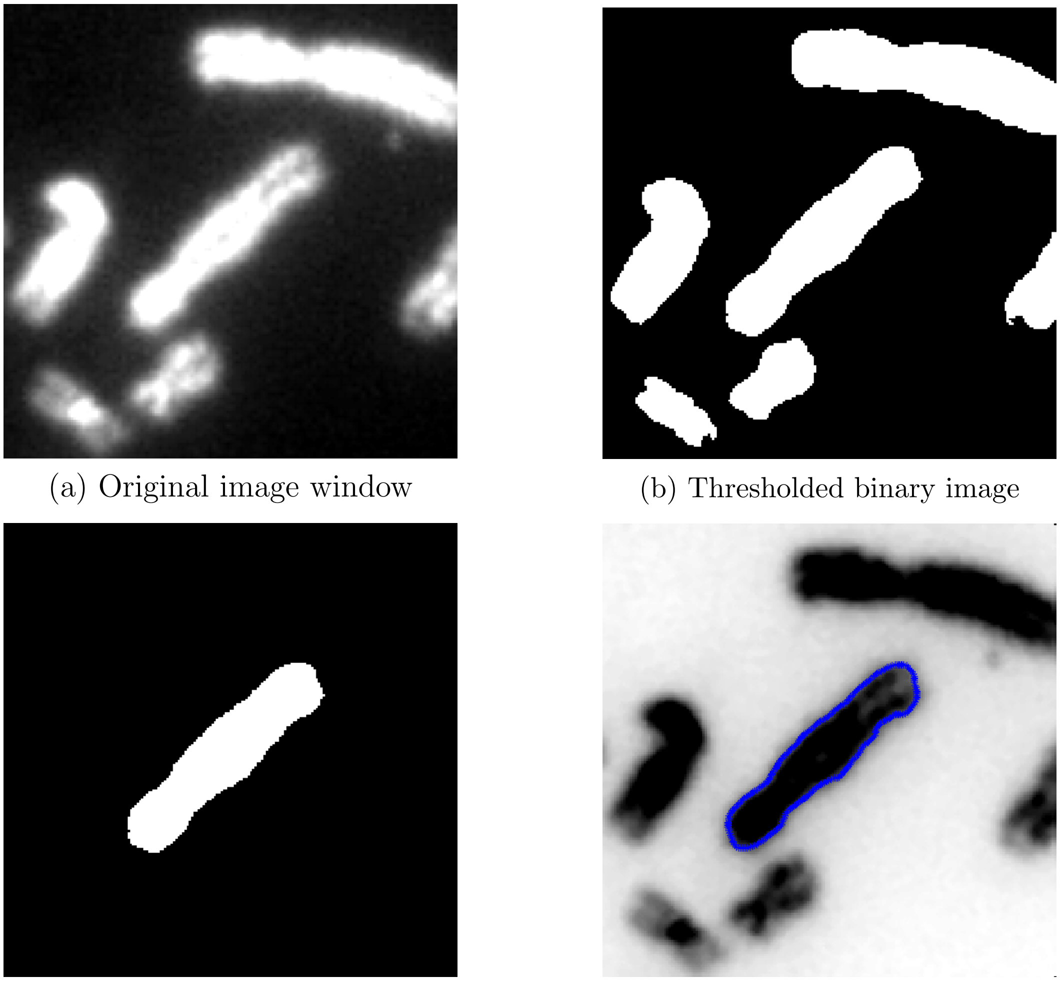
Using thresholding on an extracted window as opposed to the cell image, reduced the adverse effects of uneven illumination in cell images. However since thresholding being a point processing method, this segmentation method often yields noisy results. This can show up as both individual pixels (or small blobs) being marked as objects or as noisy object boundaries due to intensity fading in the vicinity of the chromosome boundary. The noise created by individual pixels getting thresholded as object was removed by labeling 4-connected components and then removing all regions with the same label where the regions size was less than 10 pixels (set empirically). Once these small blobs were removed, the extracted binary object was subjected to a morphological filling operation where every pixel in the image with a value '0' with 4-connected pixels with value '1' was complemented to have value '1'. Since thresholding is a point processing method, this step ensures the continuity of the chromosome blob. Next, the user selected the chromosome that needed to be extracted from the labeled binary image. Since the user selected point can be both inside, on and just outside the object boundary, the closest binary object was selected as the chromosome of interest. Next, the morphological operation was reversed on the selected chromosome blob where every pixel in the image with a value '1' which has 4-connected pixels with value '1' was complemented to have value '0'. This iterative process reveals the object boundary which was traced using a (3×3) neighborhood [47].

In order for the final stage of the segmentation algorithm to perform well, noise in a digital image needs to be removed or attenuated. Metaphase cells often have background pixels with highly variable intensity values as artifacts created during the light microscopy imaging process. A median filter with element dimensions of 5x5 was utilized in order to remove these artifacts from the ROI as a pre-processing step. Unlike Gaussian filtering, median filtering is a non linear filtering process which effectively removes noise from images without blurring object boundaries. Although the amount of noise removal is directly proportional to the size of the filtering element, it also dislocates the object boundary. Therefore a relatively smaller element (5x5) was utilized in this research. Once the image window was filtered for noise, the image window was then subjected to the next stage of the segmentation algorithm where the object boundary from global threshold was utilized as the starting points for a Gradient Vector Flow (GVF) active contour model, which is a variation of the standard parametric active contour model (see figure 3.3). The rationale for adopting a parametric active contour model was due to the availability of a close approximation of the chromosome shape through thresholding and the presence of strong edges around the chromosomes. However depending on the content within the extracted window, initial object shapes are either under-approximated or over-approximated. The GVF active contour model has the ability to both contract or expand depending on the static vector field created using the object boundary. GVF also has the capability to converge into boundary concavities which is an important property when analyzing chromosomes with high shape variability. The main internal parameters of the GVF were set at $\alpha = 0.05$ (elasticity factor), $\beta = 0$ (rigidity factor), $\mu = 0.2$ (GVF regularization factor) and $\kappa = 2$ (external force weight). This set of values were obtained empirically and yielded satisfactory segmentation results with good convergence into boundary concavities across the entire data set.

External energy models for increasing the capture range, along with the advantages of using GVF snakes are discussed in detail in section A.1. The 'Canny edge detection operator' which uses a multi-stage algorithm to accurately detect image boundary edges was utilized for generating the edge map [25]. The image result of some of the steps of the segmentation algorithm is depicted in figure 3.6.

3.3 Finding the centerline

The morphological centerline of a closed object is defined as the set of all points which are centers of circles (in 2D case) that are tangent to the shape at more than one point and that contain no other tangent circles [33]. The detection of the chromosome centerline is a necessary step in many existing chromosome analyzing methods in the literature [19], [38]. Many shape and structure-related features such as the chromosome length, chromosomal banding pattern, width and density profiles can be extracted using the centerline. Therefore the accuracy of the centerline is important since a small deviation could result in classification error [39]. The majority of the existing methods in the literature employ iterative methods such as MAT and thinning to first calculate a collection of points representing the original shape of the object [9], which is then followed by a skeleton pruning method in order to remove spurious branches. These branches are introduced due to the morphological variations of the chromosomes as well as boundary irregularities. The methods that were



(a) Original image window

(b) Thresholded binary image

(c) Extracted chromosome

(d) GVF segmentation result

Figure 3.6: Depicts the resulting image window at different stages of the segmentation algorithm where figure 3.6 (a) gives the original window containing the chromosome prior to segmentation. Figures 3.6 (b) and (c) contain the threshold output and the extracted binary object. The GVF outcome is given in figure 3.6 (d)

not based on MAT mainly have problems with handling objects with sharp bends which are commonly present in metaphase chromosomes [38], [39].

Although morphological thinning in general creates less spurious branches in comparison to skeletonization (MAT) method, both approaches are sensitive to boundary deformations, noise and irregularities and that proves to be a significant drawback in an image analysis point of view. Both methods attempt to preserve sufficient information to recreate the original shape of the object using the resulting centerline. This information includes image boundaries which are bound to be discontinuous due to quantization errors, irrespective of the resolution. This increases the probability of getting spurious branches. Figure 3.7 demonstrates some examples of chromosomes where skeletonization and thinning methods create spurious branches. Due to the highly unpredictable nature of the appearance, these branches can pose a significant difficulty for any higher level image analysis algorithm thus demanding a reliable skeletal pruning algorithm in order to obtain the centerline. The majority of skeletal pruning algorithms reported in the literature are application specific and are based on simple methodologies. The Prairie fire model [33] is one of the most common methods utilized in which the propagation velocities were adjusted to be proportional to the curvature at the fire front and by doing so attempts to promote convexity of the binary object during skeletonization. Another method attempted to remove branches with a shape contribution below a certain threshold level in order to prune these spurious branches [48]. However these methods have the tendency to provide disconnected skeletal segments due to the lack of regularization.

In this research a skeleton pruning method based on Discrete Curve Evolution (DCE) [49] was applied to obtain an accurate centerline of a chromosome with any morphological variation having no spurious branches. This algorithm first partitioned the contour into a given number of polygonal sections which in turn was then used for removing all spurious branches in the skeleton. The partitioning was obtained by evolving the polygonal structure using DCE. Furthermore, pruning was achieved by removing all skeletal points of which all the generating points (the points where the maximal disks touch the object boundary) lie on the same polygon partition. Results in this skeletal pruning method were highly dependent on the contour partitioning itself. Therefore the skeleton pruning problem was transformed to a problem of obtaining an accurate contour partitioning which represents the original shape.

DCE was the ideal solution for this problem where any shape can be simplified

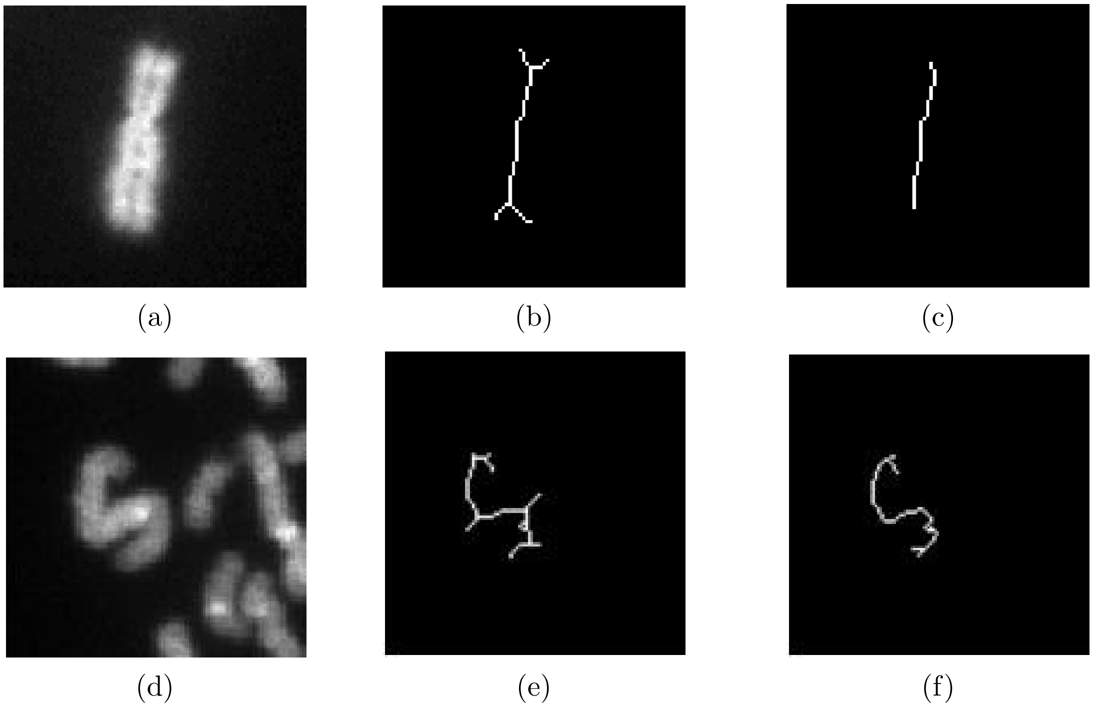


Figure 3.7: Two chromosomes and their skeletonization [3] and morphological thinning results [4] showing some resulting spurious branches. These operations were performed on the binary object obtained through Otsu's method. Figure 3.7 (b) and (e) are the skeletonization results while figure 3.7 (c) and (f) are the morphological thinning results.

by effectively evolving polygon partitions by vertex deletion based on any given relevance measurement [50]. Since any digital image boundary is easily approximated to a polygon without loss of information by taking each boundary pixel as a vertex on the polygon where the distance between each of these can be considered as the edges. DCE was then used to evolve the polygon iteratively by removing the vertex which had the least value for the relevance value $K(v, u, w)$ defined in equation 3.3, where d_{uv} & d_{vw} are the Euclidean length between the vertices and θ is the turn angle at vertex v. This relevance function was selected so that it is dependent on features of its neighbors thus making DCE able to evolve using global features of the shape information.

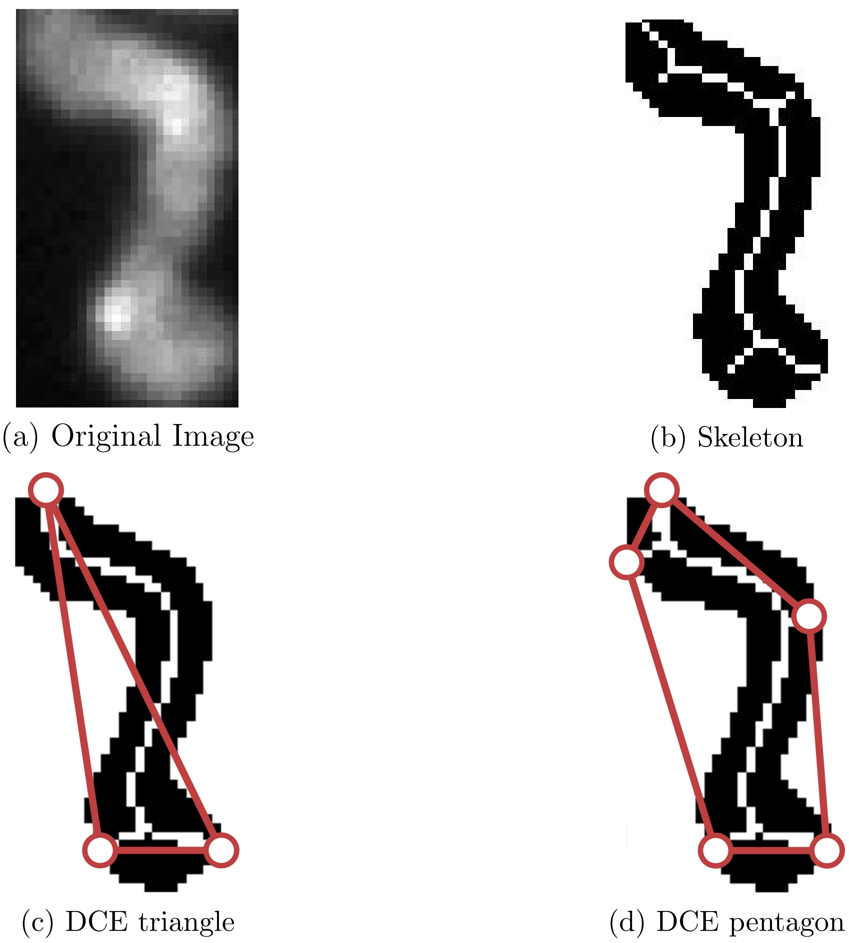
$$K(v, u, w) = (\theta * d_{uv} * d_{vw})/(d_{uv} + d_{vw}) \tag{3.3}$$

Topology information of the original shape is guaranteed to be preserved at the skeleton ends since the evolution proceeds by deleting vertices as opposed to methods that displace the vertices of the polygon. Furthermore, the speed of the iterative evolution process was improved by deleting multiple vertices which share the same lowest relevance value at the initial stage. The DCE algorithm is also highly robust against noise on the object boundary since such points are deleted at the early stages of the evolution due to their relatively smaller relevance measure with respect to the original shape of the object. Another important property of the DCE method is that the algorithm can guarantee convergence, which will be the polygon with the number of vertices set as the end stop criteria.

The telomere regions where the ends of the centerline should fall are convex in shape. Therefore by considering only convex polygon combinations, the possibility of selecting a bending point of a chromosome (which is concave) as an end point for the centerline was significantly reduced [6], [49].

Figure 3.8 depicts the polygonal skeletal results at various stages of the DCE based skeletal pruning algorithm compared to the standard skeletonization result. Figure 3.8(c) and figure 3.8(d) are the skeletons resulting from the DCE triangle and the DCE pentagon.

In obtaining the medial axis of a chromosome, the ideal result would be a pruned skeleton with no extra branches. However, since the minimum convex polygon being a triangle and DCE being modeled as convex polygons, the resulting skeleton will have one spurious branch. Although this method leaves a spurious branch in


(a) Original Image

(b) Skeleton

(c) DCE triangle

(d) DCE pentagon

Figure 3.8: Comparison between standard skeleton with DCE based skeletal pruning results where (c) and (d) are the skeletons resulting from the DCE triangle and the DCE pentagon results.

the skeleton, it is a consistent occurrence as opposed to the unpredictable nature involved in other morphological approaches. Throughout this chapter, the symbol P will refer to various point sets on the chromosome object contour. If $C \in \mathbb{R}^2$ is the contour of the chromosome, the DCE initial anchor points (skeletal end points) for the centerline are denoted by $P^{\hat{E}}$ ($\left|P^{\hat{E}}\right| = 3$). Since the DCE method preserves the topological information of the chromosome, the spurious branch is simply removed by tracing all branches and pruning the shortest branch completely. This yields the set of anchor points P^E. This set of points is be used for partitioning the chromosome contour in order to isolate the telomere region in section 3.4.1.

The DCE result was then processed by a modified thinning algorithm to ensure

-1	-1	+1
-1	+1	+1
-1	-1	0

(a)

-1	-1	0
-1	+1	+1
-1	-1	+1

(b)

-1	-1	-1
-1	+1	-1
0	+1	+1

(d)

-1	-1	-1
-1	+1	-1
+1	+1	0

(e)

+1	-1	-1
+1	+1	-1
0	-1	-1

(a)

0	-1	-1
+1	+1	-1
+1	-1	-1

(b)

+1	+1	0
-1	+1	-1
-1	-1	-1

(d)

0	+1	+1
-1	+1	-1
-1	-1	-1

(e)

Figure 3.9: The eight masks used for the hit and miss algorithm where '+1' and '-1' elements search for matching foreground and background pixels respectively, while the '0' ignores the value at that position.

single pixel thickness of the skeleton. This modified thinning algorithm consisted of the application of a set of masks to the skeleton on the basis of the morphological hit & miss algorithm followed by a thinning process described by Lam [4]. The hit and miss algorithm was applied prior to the application of the standard thinning method for ensuring that the junction points where two branches connect have a single pixel thickness. Figure 3.9 depicts the set of eight masks used for this step where '+1' and '-1' elements search for matching foreground and background pixels respectively, while the '0' ignores the value at that position.

The skeletonisation result was next pruned by 10% (empirically set) of the total length of the skeleton at both ends. This accounted for the skeletal bifurcations where the skeletal portion deviated at the telomere regions from the actual centerline. It is important to note that this pruning percentile value is not universal. However, since the proposed method does not use the centerline of the chromosome as a basis for measurements, the accuracy of the centerline is not detrimental to centromere localization. Finally, a simple method was devised to correct the end points of the pruned centerline using the GVF segmentation outcome. End points of the pruned centerline were used with 2 points sampled 7 pixels inwards from both ends to calculate the inclination. This slope was then used to extend the ends of the pruned centerline until the GVF segmentation outline was intersected.

3.3.1 Centerline based Centromere identification

Apart from calculating the chromosome length, the centerline provides a basis for detecting centromere locations as well. This is a rational step since the centerline derived in the above algorithm is not susceptible to spurious branches. Since the telomere regions of the chromosome are narrow compared to the body of the chromosome, the pruned centerline was used as the basis for calculating the width profile, which is the collection of width measurements along the longitudinal axis of symmetry. Width constriction is a characteristic feature of a chromosome and therefore the width profile is the best feature for detecting the centromere location.

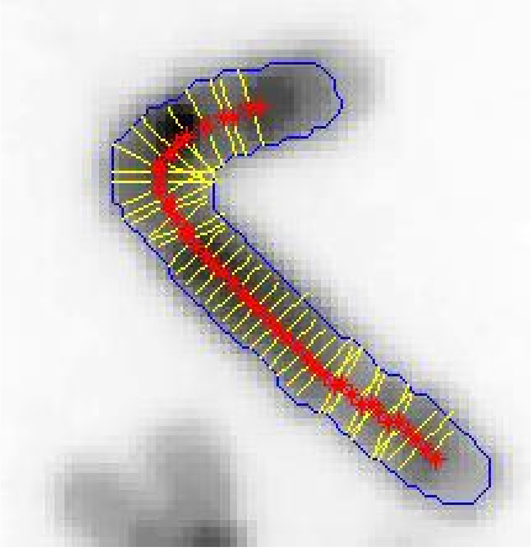


Figure 3.10: Depicts the trellis structure created for calculating the width profile of a chromosome.

The width profile was calculated using a set of perpendicular line segments(called the 'trellis structure') along the centerline. Figure 3.10 demonstrate such a trellis structure which also can be used for detecting other features such as intensity feature profiles of the chromosome. Once the width profile was generated, the global minima of this profile was selected as the chromosome centromere location. This simple approach yielded reasonably accurate results. However, the presence of sister chromatid separation was observed to dislocate the centerline by placing it on one of the sister chromatids as opposed to the middle of the chromosome. This created false measurements for the width profile where the width measurements were only based on one of the chromatid arms, which leads to false localization on one of the chromatids. Figure 3.11 depicts examples of such cases where the centromere localization algorithm was mislead by the presence of this artifact. The algorithm discussed in the subsequent sections has the capability to detect and correct for sister chromatid separation and therefore mitigate the impact of the artifact on the detection accuracy.

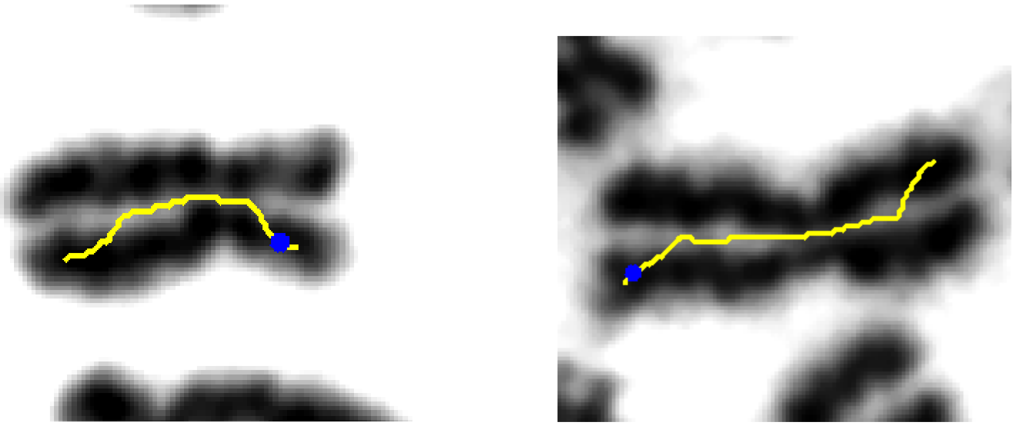

(a) (b)

Figure 3.11: (a) and (b) depict two instances where the centerline based centromere detection was adversely affected by sister chromatid separation.

3.4 Contour partitioning & correcting for sister chromatid separation

Sister chromatid separation in chromosomes is an integral process in cell division or mitosis. Therefore depending on the stage of mitosis at which the cells were arrested, sister chromatid separation may be visible at a varying degree. Furthermore, a chemical agent termed colcemid which is used mainly as a preparatory chemical in biodosimetry studies, can cause or exacerbate this condition and prematurely force sister chromatid separation on metaphase cells. Therefore it is important that a given chromosome processing software be able to analyze chromosomes with sister chromatid separation to the greatest possible extent. However from an image analysis point of view, the presence of sister chromatid separation significantly increases the complexity of the morphological variations of the chromosome (see figure 3.13). The most substantial change of shape features occur towards the telomere regions

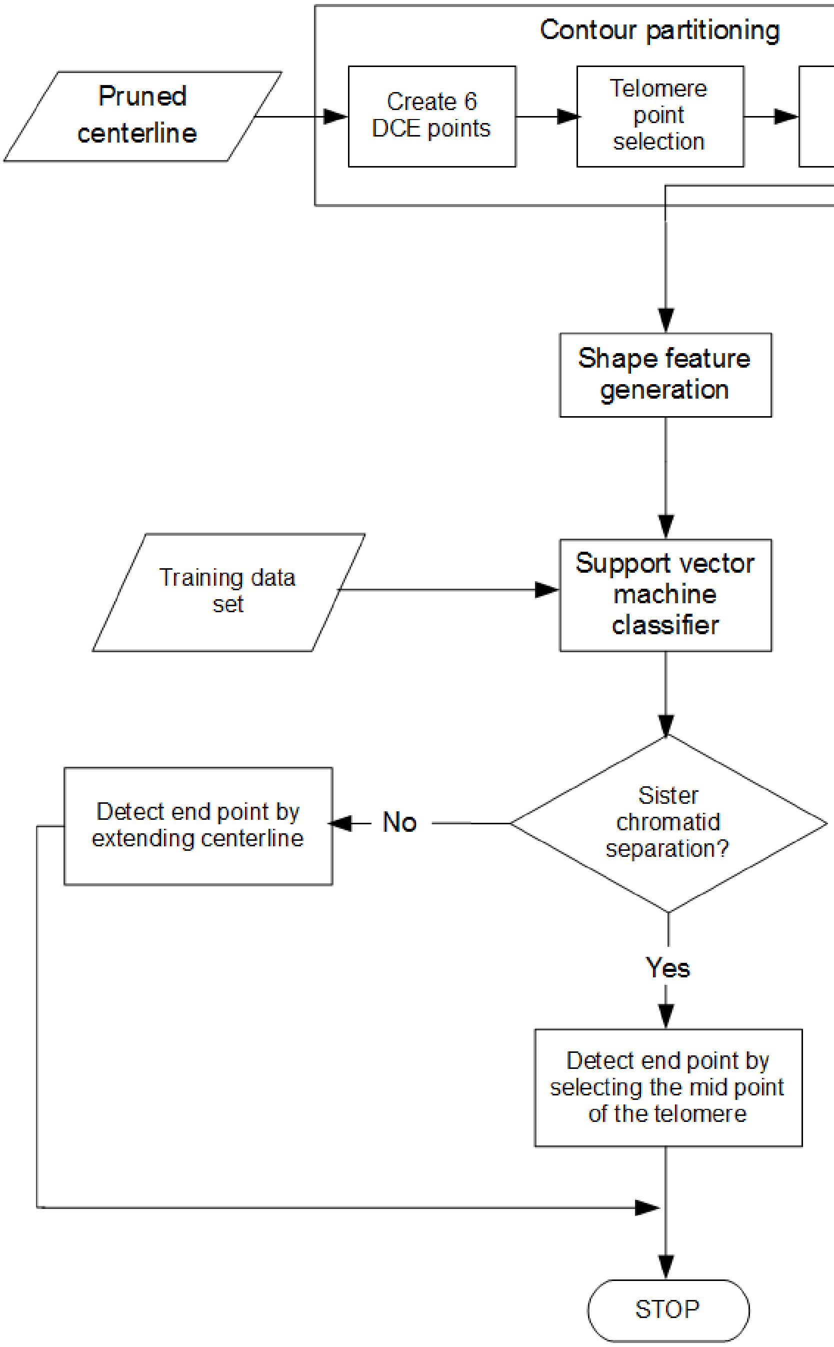

Contour partitioning
Pruned centerline
Create 6 DCE points
Telomere point selection
Partition telomere region
Shape feature generation
Training data set
Support vector machine classifier
Sister chromatid separation?
No
Detect end point by extending centerline
Yes
Detect end point by selecting the mid point of the telomere
STOP

Figure 3.12: The flow diagram of the correction for sister chromatid separation stage of the proposed method.

of the chromosome where the separation of the sister chromatids create concavities in the segmentation outline. These concavities attract the centerline into one of the sister chromatids, leading to false localization of centromere in one of the chromatid arms. Therefore it is essential to detect and correct for sister chromatid separation in metaphase cell images in order to develop a reliable chromosome analysis algorithm. The following sub section describes the automated contour partitioning and shape matching algorithm that has been proposed in this research to identify and correct for sister chromatid separation. The steps involved in detecting and correcting for sister chromatid separation is given by figure 3.12.

Figure 3.13: Depicts the effects on the chromosome morphology introduced by sister chromatid separation. (a) and (b) depict two straight chromosomes without and with sister chromatid separation respectively both with DAPI staining.

The chromosome thickness measurement algorithm discussed in section 3.5 requires an approximately symmetric division of the contour of the chromosome. Accurate partitioning of the telomere region can yield means to identify evidence of sister chromatid separation and therefore correct for any such artifact as well as to split the contour into two segments accurately.

3.4.1 Contour partitioning for isolating the telomere

Contour partitioning requires an algorithm which can effectively detect the salient points despite the morphology of the chromosome. Curvature of the contour is one of the most commonly used features in the literature for detecting salient points that can be used for partitioning chromosomes [51]. An important requirement is that the location of these salient points need to be highly repeatable under varying level of object boundary noise. Curvature values of the chromosome boundary are highly susceptible to the boundary noise which is introduced by different staining and cell preparation techniques.

The DCE method used for polygon evolution and shape simplification for pruning spurious branches in the centerline is immune to boundary noise. This is because the noisy boundary points are deleted during the initial stages of the DCE algorithm owing to the lack of contribution of those points to the overall shape of the object. Furthermore the DCE method also preserves the topological information of the original object during shape simplification. These properties provide an ideal platform for obtaining a set of initial salient points on the contour of the chromosome outline which performs well with boundaries regardless of their smoothness, yielding repeatable results [52]. The ability to terminate the process of DCE shape evolution at a given number of vertices further lends to its applicability. The requirement of the contour partitioning stage was to detect the 4 salient points of the chromosome which isolates the two telomere regions of the chromosome. However, due to morphological variations, the DCE result with just 4 contour points (DCE rectangle) could not guarantee the inclusion of the points required for isolating the telomere region. Next it was empirically established that a termination at 6 DCE points would ensure that the required telomere end points will be retained within the set of candidate salient points. However, the two anchor points (P^E) of the skeleton obtained through DCE based skeleton pruning in section 3.3 is a subset of these 6 DCE points. Contour partitioning was performed by selecting the best 4 point combination (including the two anchor points) that represents all the telomere end points.

The approach for selecting the best contour partitioning combination has two stages as listed below,

- Training a feature based classifier using a large data set to capture desirable properties which contribute to make a good contour partitioning combination.
- Using the trained classifier for selecting the best combination from a set of contour partitioning set of points.

At the first stage, all the combinations across the data set were used as a pool of candidates to train the classifier. Use of this data aided the classifier to capture the desirable features across the data set and account for the high morphological variations in human metaphase chromosomes. Once trained, the classifier was then used to place the 12 candidates of each chromosome in the feature space separately. Then the signed Euclidean distance from the separating hyperplane (say ρ) was calculated for each of the candidate for the given chromosome. This distance was directly used as a measure of the goodness of fit of a given set of combinations. The best combination for partitioning the chromosome contour for isolating the telomere region was selected by picking the candidate with the largest distance from the separating hyperplane. Unlike traditional rule based ranking algorithms, this approach required very little high level knowledge of the desirable characteristics. The positioning of the separating hyperplane encapsulated this high level information through user-provided ground truth. Therefore the highest ranked combination was the best set of points that could be used to isolate the telomere region to detect the evidence of sister chromatid separation.

In order to define the features (F^s) used for contour partitioning, Let Φ_h be the curvature value at candidate point h and $S \in \mathbb{R}^2$ be the skeleton of the chromosome with 6 DCE point stop criteria. Next, the following set of points were defined,

- P^D ($\subset C$) is the set of six DCE vertices.
- P^S constitutes of all the points in P^D except the anchor points (P^E). These are the four telomere end-point candidates.

Then the family of sets P^T for all possible combinations with the sets P^E and P^S would contain the following combinations,

$$\left\{P_1^E, P_1^S, P_2^E, P_2^S\right\}, \left\{P_1^E, P_1^S, P_2^E, P_3^S\right\}, \left\{P_1^E, P_1^S, P_2^E, P_4^S\right\},$$
$$\left\{P_1^E, P_2^S, P_2^E, P_1^S\right\}, \left\{P_1^E, P_2^S, P_2^E, P_3^S\right\}, \left\{P_1^E, P_2^S, P_2^E, P_4^S\right\},$$
$$\left\{P_1^E, P_3^S, P_2^E, P_1^S\right\}, \left\{P_1^E, P_3^S, P_2^E, P_2^S\right\}, \left\{P_1^E, P_3^S, P_2^E, P_4^S\right\},$$
$$\left\{P_1^E, P_4^S, P_2^E, P_1^S\right\}, \left\{P_1^E, P_4^S, P_2^E, P_2^S\right\}, \left\{P_1^E, P_4^S, P_2^E, P_3^S\right\}.$$

Figure 3.14 illustrates one such combination where the selected (connected by the blue line segments) combination for the contour partitioning points are given by $\left\{P_1^E, P_4^S, P_2^E, P_1^S\right\}$.

The SVM classifier utilized for detecting the best combination for contour partitioning was trained using 11 features (F^s) listed subsequently. A Gaussian radial basis function kernel was used with sequential minimum optimization (which yields a l-1 norm soft margin classifier) for training this classifier. These were designed to capture a collection of features capturing local properties of each point and measure of how they position relatively as well. Features F_1^s and F_2^s provide an indication to the saliency of each of the candidate points to decide whether the candidate point was a skeletal end point during the pruning process. This provided a measure of saliency with respect to the morphological skeletonisation process. Features F_3^s to F_5^s are three normalized features which capture the relative positioning of each candidate in the given combination. F_6^s and F_7^s represent the shape or the morphology of the chromosome of interest. Although the value of F_6^s and F_7^s is the same for all 12 candidates for given chromosome, the inclusion of these features account for morphological variations across the cell images in the data set. F_8^s and F_9^s represent the curvature of the candidate points as well as the concavity/convexity of those locations. This is an important feature since salient points used for contour partitioning are in general convex. The features F_{10}^s and F_{11}^s are two Euclidean distance based features which captures the proportion of each telomere region in the combination to the perimeter of the rectangle made by connecting the 4 points in consideration. During the research, a significant improvement of the accuracy of classification was observed by the inclusion of these two features.

Let $d\,(p,q)$ denote the Euclidean distance between the points p and q. Similarly let $l\,(p,q)$ represent the length of the curve between p and q, which are points from the set P^D. Then for each contour partitioning combination in P^T given by $\left\{P_1^E, P_i^S, P_2^E, P_j^S\right\}$ (where i and j are integer values such that $1 \leq i,j \leq 4$ and $i \neq j$), two main length measurements ratios (r_1 and r_2) are used for both calculating length based features as well as for normalizing them. $r_1 = \frac{l(P_1^E,P_i^S)}{l(P_1^E,P_j^S)}$ which yields the chromosome width/length with respect to the anchor point P_1^E for the given contour partitioning combination (refer figure 3.14). Similarly $r_2 = \frac{l(P_2^E,P_i^S)}{l(P_2^E,P_j^S)}$ is calculated with respect to the anchor point P_2^E. Then the set of features F^s for each contour partitioning combination is defined as follows,

1. $F_1^s = 1$ if the point P_i^S belongs to a skeletal end point ($P_i^S \in (S \cap C)$). Otherwise, $F_1^s = 0$.

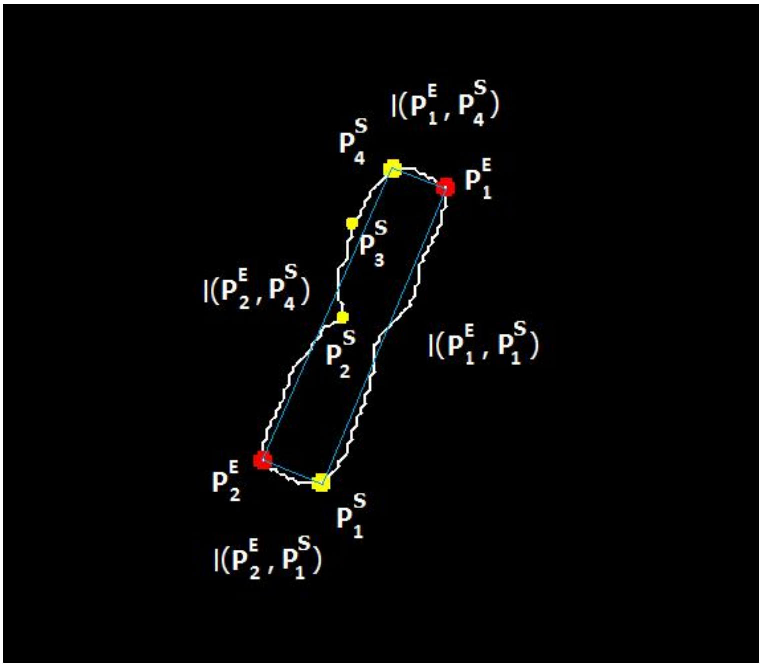


Figure 3.14: Figure demonstrates one possible combination for contour partitioning where the anchor point (red '+' sign) P_1^E is connected with the candidate point P_4^S while the other anchor point P_2^E is connected with candidate point P_1^S which captures the telomere regions. The blue line connects the set of points constituting the considered combination in this instance.

2. $F_2^s = 1$ if the point P_j^S belongs to a skeletal end point ($P_j^S \in (S \cap C)$). Otherwise, $F_2^s = 0$.

3. $F_3^s = \left[1 - \left|\frac{r_1 - r_2}{max(r_1,r_2)}\right|\right]$ where $0 < F_3^s < 1$. This calculates the chromosome width/length ratio for each anchor point and the difference between the two measures. Two similar fractions would result in a high value for the feature F_3^s.

4. $F_4^s = \left[1 - \frac{r_1}{max(r_1,r_2)}\right]$ where $0 < F_4^s < 1$. This calculates the chromosome width/length ratio with respect to the first anchor point (P_1^E). The telomere region in general is shorter than the sides of the chromosome. Therefore a lower length ratio measurement which in turns is a higher value for the feature F_4^s is a desirable property.

5. $F_5^s = \left[1 - \frac{r_2}{max(r_1,r_2)}\right]$ where $0 < F_5^s < 1$. This is same as F_4^s, but from the other anchor point, P_2^E.

6. F_6^s : ratio of length of the chromosome to area of the chromosome. This provides a measure of elongation of a chromosome.

7. F_7^s : ratio value of perimeter of the chromosome to the area of the chromosome. This provides a measure of how noisy the object boundaries are.

8. F_8^s : average of the curvature values Φ_h of the candidates. The curvature is an important measurement of the saliency of the candidate points.

9. F_9^s : number of the negative curvature values ($\Phi_h < 0$) of the candidates points $\left(P_i^S \, and \, P_j^S\right)$. The telomere region end points are generally characterized by points with high convexity. The number of negative angles yield how concave the points of interest are.

10. $F_{10}^s = \frac{d(P_1^E, P_i^S)}{D}$ where $D = \sum_{x=1,y=i,j}^{x=2} d(P_x^E, P_y^S)$. This feature calculates the normalized Euclidean distance between the anchor point P_1^E and the candidate P_i^S which makes up one telomere region.

11. $F_{11}^s = \frac{d(P_2^E, P_j^S)}{D}$ where $D = \sum_{x=1,y=i,j}^{x=2} d(P_x^E, P_y^S)$. Same as feature F_{10}^s but calculated for the other anchor point.

Once the set of features were finalized, the combinations for the contour partitioning set of points were created for the complete data set. This created 16,800 combinations for the total of 1400 metaphase chromosomes. Next, ground truth required for training and testing the accuracy of the classifier was collected. The author examined all the possible combinations for each chromosome in the data set and manually marked the combinations which are viable solutions for partitioning and isolating the telomere regions of the chromosome. Next, a SVM classifier was trained and tested for effectiveness with 2 fold cross validation (50% - train data, 50% - test data) and obtained an accuracy, sensitivity and specificity values of 94%, 97% and 68% respectively. The results demonstrated the ability of the feature set to effectively detect good combination of candidate points for partitioning telomere regions. Slightly lower specificity suggests inclusion of some false positives into the detection. However, this does not affect the accuracy of the contour partitioning since the algorithm picks the combination based on its rank as opposed to the classification label.

Once the 12 combinations for a given chromosome was generated, the signed distance from the separating hyperplane (ρ) was calculated. This distance is the geometric margin of that sample in the feature space with respect to the separating hyperplane (see section A.3). Then the combination of points with the largest positive geometric margin (ρ) was selected and the telomere regions was segmented using these points. The algorithm used for the analysis of shape features of these telomere contour segments is described in section 3.4.2.

3.4.2 Shape information extraction

The partitioned telomere curve segments carry important information relating to the presence of premature sister chromatid separation. Traditionally, this is achieved by incorporating high level shape information using a large set of well defined features which capture the subtle changes in shape characteristics. Once a large number of features have been defined, it is common to try and reduce the number of features to optimize the feature space for classification. Principal component analysis (PCA) is one of the most common methods adopted to reduce the number of features. However, there are few disadvantages in a feature based set up. Namely,

- The performance of the system directly depends on the defined set of features. Therefore, this will lead to experimentation with a variety of feature combinations in order to optimize the setup.

- Some shape features are not invariant to the scale of the object. In these cases, additional computing is required to make the features scale invariant.

- The lack of expert knowledge of how each of the features contribute towards detecting the required shape.

In this research, an alternative approach towards detecting shape features was proposed using functional approximations of each of the curve partitions as an alternative to various geometrical features. The concept of using orthogonal coefficients for matching shapes have been tested in the field of hand writing recognition with satisfactory results [53], [54]. The procedure of the algorithm in calculating the coefficients (based on Legendre polynomials) is explained below.

Orthogonal function representation:

By definition, a set of functions is considered 'Orthogonal' with respect to a functional inner product $< . , . >$ within the domain $[\, a, b \,]$ if,

$$< h_i, \, h_j > = \int_a^b h_i(t).h_j(t).w(t)\, dt \; = \; 0 \;\;, \, i \neq j \tag{3.4}$$

where, $w(t)$ is the weight function defined for the same domain. A given function $f(t)$ is expressed as a linear combination of coefficients calculated related to an orthogonal basis (refer equation 3.5).

$$f(t) \; = \; \sum_{i=0}^{\infty} \alpha_i.h_i(t) \tag{3.5}$$

Where each coefficient α_i can be calculated as follows,

$$\alpha_i \; = \; \frac{< f, h_i >}{\|h_i\|^2} \tag{3.6}$$

However, the functions of the contour segments are unknown at the moment. Therefore equation 3.6 cannot be directly used to calculate the coefficient values. Instead, we need to calculate moments of different orders using these coordinate values (for x and y coordinates separately). By definition, the k^{th} order moment of the x coordinate is given by,

$$\mu_k \; = \; \int_a^b x^k f(x) dx \tag{3.7}$$

The coordinate points in the partitioned contour segments are not uniformly distributed. This is due to the forces acting upon each control point in the GVF segmentation stage. Therefore, each contour segment has to be parameterized accordingly. Since the curve segment length is variable, coordinate points have to be parameterized using the arc length λ to obtain $f(\lambda)$. Given the x and y coordinates, the Euclidean distance between each coordinate pair ($dx1$,$dy1$) were first calculated, along with the cumulative distances ($d1$) and offsets ($f_x(\lambda)$,$f_y(\lambda)$) from the initial position. With this information, the moment of a given order within two consecutive contour points using equation 3.8 was calculated. Then, each moment (μ_k) for the respective variable is calculated by taking sum of the moments for all intervals (refer equation 3.8).

$$\begin{aligned}\int_{i}^{i+1} \lambda^k f(\lambda) d\lambda &= \frac{d1(i+1)^{k+1} - d(i)^{k+1}}{k+1} \\ &\times \frac{f(i+1)+f(i)}{2} \qquad (3.8)\end{aligned}$$

In this research the series of moments were truncated at the ninth order element. Therefore it is assumed that calculation of ten moments accurately recreate or represent a shape characteristics at the telomere region. This is a reasonable threshold since the inclusion of moments of higher order values adds very little information of the general shape while having the possibility of including boundary noise of the chromosome outline [54]. Next, a set of orthogonal basis functions were created using Legendre polynomials ($P_n(t)$) given by equation 3.9 with a weight function $w(t) = 1$, which are orthogonal in the interval $[-1, 1]$. These polynomials are easily generated using a symbolic mathematical software such as MAPLE [55] or Mathematica [56].

$$P_n(t) = \frac{1}{2^n n!} \frac{d^n}{dt^n} (t^2 - 1)^n \qquad (3.9)$$

Next, the series coefficient were calculated using the moment values. The calculation of these coefficients was performed using equation 3.10, where k is the maximum order of the moments and L is the total length of the curve segment. These coefficient values were scaled in order to span any required interval $[a, b]$ using equation [53], where $\delta_{i,0} = 1$ only when $i = 0$.

$$\begin{aligned}\hat{\alpha}_k &= (-1)^k \frac{2k+1}{L} \\ &\times \sum_{i=0}^{k} (\frac{-1}{L})^i \binom{k}{i} \binom{k+i}{i} \mu_i \qquad (3.10)\end{aligned}$$

$$\hat{\hat{\alpha}}_i = \hat{\alpha}_i \frac{b-a}{f_{max} - f_{min}} + \delta_{i,0} \frac{a.f_{max} - b.f_{min}}{f_{max} - f_{min}} \tag{3.11}$$

With the imposed limitation of 10 coefficients per axis, this yielded 20 features representing shape features along both x and y axis. A second SVM classifier was trained using 90 labeled set of telomere coordinate curve segments to effectively detect shape variations inherent to telomere regions with sister chromatid separation. A multi-layer perceptron kernel was used with Quadratic programming (which yields a l-2 norm soft margin classifier) was was utilized for training this classifier. These telomere curve segments were examined for evidence of sister chromatid separation which appears as a boundary concavity and was manually labeled for training and testing the classifier. The support vector machine classifier is a 'large margin classifier', which maximizes the largest distance to the nearest training data points of any class [57]. This yielded more reliable classification for any new data points even close to the decision boundary/plane. With 2-fold cross-validation (50% - train data, 50% - test data), the support classifier demonstrated an accuracy higher than 92%. If a telomere region was detected for sister chromatid separation, then the end point correction discussed in section 3.3 was altered so that the extended line satisfy the coordinates of the telomere mid point (see figure 4.2). This correction was not meant for correcting the centerline of the chromosome for the artifact of sister chromatid separation. Instead it attempts to split the contour of the chromosome into two approximately symmetrical contour segments which is a requirement for the Laplacian based thickness measurement algorithm proposed in section 3.5 below.

3.5 Laplacian based thickness measurement

An effective framework for calculating the width profile of a human metaphase chromosome should posses the following properties,

- The effect of boundary noise needs to be minimum since it introduces noisy measurements in the width profile values. Noise on the object boundary gets represented in the centerline. The scan lines calculated based on centerline tend to miss the actual constriction at the centromere location. This can lead to high false positives in the centromere localization process.

- The measurement requires to be uniform in sampling the width of the chromosome along the longitudinal axis of the chromosome. This is specially important at bends of the chromosome since most bends happen at the centromere of the chromosome.

All the centerline based measurement methods, including the trellis structure method described in section 3.3.1 are prone to creating incorrect width measurements due to noisy centerline data points introduced through boundary noise. Furthermore, correcting the centerline for sister chromatid separation is a difficult problem as well. Therefore, a better algorithm was proposed which uses the centerline merely for dividing the chromosome into two sections and not as a basis for width or intensity profile measurements. This algorithm was based on solving the Laplacian heat equation for a contour image and was further modified to include intensity information into the process.

3.5.1 Intensity integrated Laplacian based thickness measurement

Laplacian based thickness measurement is an algorithm satisfactorily used for cortical thickness measurements in some brain mapping applications [58], [59]. The Laplacian operator (Δ) yields the divergence of the gradient of a function in the Euclidean point space. In other words, it gives the difference of the gradient or the second order derivative of a function or an image. This is written as follows, where $\bigtriangledown$ is the first derivative or the divergence of the gradient operator in any given direction.

$$\Delta f \;=\; div(\bigtriangledown f) \;=\; \bigtriangledown . \bigtriangledown\, f \tag{3.12}$$

In this approach the Laplacian operator was used to obtain the steady state of heat flow or voltage distribution between two heated or charged surfaces. This stage of the algorithm operated on the contour of the chromosome which was split into two approximately symmetrical segments after correcting for sister chromatid separation in section 3.4.2. By retaining the two longitudinal contour sides at two different potentials or temperatures, a set of equipotential lines in the static vector field created by the heat flow was derived in steady state according to the Laplacian equation [59]. Depending on the relative voltage or temperature that the segments

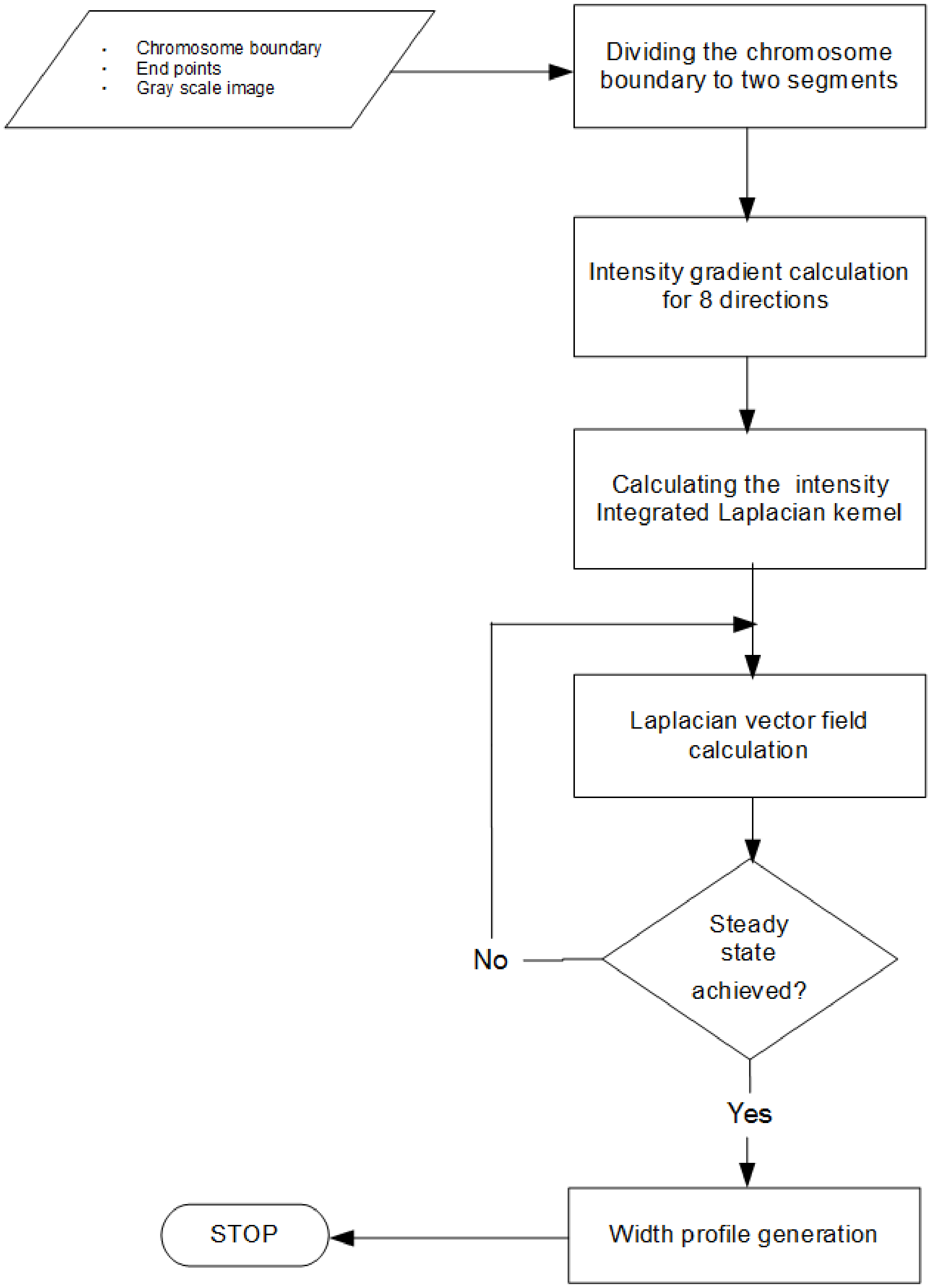

Chromosome boundary
End points
Gray scale image
Dividing the chromosome boundary to two segments
Intensity gradient calculation for 8 directions
Calculating the intensity Integrated Laplacian kernel
Laplacian vector field calculation
Steady state achieved?
No
Yes
Width profile generation
STOP

Figure 3.15: The flow diagram of the Intensity integrated Laplacian algorithm of the proposed method.

were maintained, a heat flow static field was initiated from one segment and end on the other segment creating a path. Therefore a simple incremental method such as the Euler's method was utilized to trace the thickness of the chromosome from one contour segment to the other by traversing normal to these equipotential lines [60]. Then the thickness or width was calculated by summing up the Euclidean length of all the small incremental segments together. This method gives a uniform sampling of the width profile better than other techniques based on the centerline.

However, due to the sole dependency on the contour information, the Laplacian based method can still be susceptible to contour noise embedded during the segmentation stage. With different staining methods and imaging conditions, the object boundary noise content may vary significantly and in return will affect the effectiveness of this algorithm. On the other hand, most chromosome images contain some amount of intensity band information. The amount of visibility of this banding information varies across different staining and cell preparatory techniques. The direction of the intensity information is also useful since they are in general oriented normal to the object contour. Furthermore, the centromere region in general has a homogenous intensity patch surrounding the location. Therefore, intensity information carries useful information that can assist the thickness measurement process. This creates the need for a framework for integrating this valuable intensity information into the thickness measurement process and thereby reducing the influence of boundary noise on the Laplacian based algorithm. The solution proposed was to integrate intensity information using a simple weighting scheme which captures both the direction and the magnitude of intensity values in the neighborhood. This information was then used to improve the thickness measurement algorithm as discussed below. The main objective of the inclusion of the intensity is to guide the Laplacian static field across the breadth of the object, based on neighboring pixel intensity values. Figure 3.15 provides a flow diagram depicting the main stages of the intensity integrated Laplacian method.

The proposed thickness measuring algorithm required the following information as inputs to the system.

- The single pixel wide contour of the segmented object of interest.
- A separation of the object contour using the longitudinal axis of symmetry of the object. Correcting this for sister chromatid separation was performed

through the shape analysis process.

When applied to digital images, the standard Laplacian equation is represented by the kernel given in table 3.1. This digital representation makes the application of the Laplacian method simpler on high resolution images. The constant scaling factors in the kernel given in table 3.1 implies that the algorithm affects the $3x3$ neighborhood uniformly based on the contour information. The proposed weighting scheme for including intensity information simply changes the Laplacian kernel weights to account for local intensity variations as described below.

Table 3.1: A kernel that represents the Laplacian equation on a digital image

$-\frac{1}{8}$	$-\frac{1}{8}$	$-\frac{1}{8}$
$-\frac{1}{8}$	$+1$	$-\frac{1}{8}$
$-\frac{1}{8}$	$-\frac{1}{8}$	$-\frac{1}{8}$

Given the intensity image (I) which contains the object of interest, a total of 8 matrices (digital images) were created based on connectivity and directional intensity gradients with identical dimensions to I as follows,

$$\begin{aligned} \nabla \mathbf{I}_{(i,j)} &= abs[I(x,y) - I(x+i, y-j)] \\ (i,j) &= \{i, j \in (-1,0,1), (i,j) \neq (0,0)\} \end{aligned} \tag{3.13}$$

For simplicity and clarity, remaining steps will be described using the generic term $\nabla \mathbf{I}_{(i,j)}$. These 8 gradient images capture the intensity variations in all the 8 major directions on the $3X3$ neighborhood of each pixel in the image yielding a higher value in the directions across the edges of the chromosomal bands. Due to the need of capturing subtle intensity variations to bias the vector field, the above crude method of calculating the gradient was preferred over methods such as Canny edge detection. Next, all the matrices were normalized to the interval $(0\,,\,1)$, using the maximum absolute intensity difference in that direction (refer equation 3.14). Then, the matrix values were inverted within the same range of (0 , 1) by subtracting each matrix value from 1. The matrix $\nabla\mathbf{I}_{(i,j)}$ then yielded values close to unity where intensity level in the neighborhood was similar. Similarly this also gave smaller values (close to 0) for pixels with high intensity gradients. To address cases where intensity patches were parallel to the object contour, the proposed algorithm can be modified by simply removing the inverting step for all 8 matrices. By doing so, the weighting factors will bias towards higher intensity differences instead of homogenous regions. This is a useful capability when adopting this algorithm into other fields where the banding information orientation is different from that of human metaphase chromosomes.

$$\nabla\mathbf{I}_{(i,j)} \;=\; \frac{\nabla\mathbf{I}_{(i,j)}}{max(\nabla\mathbf{I}_{(i,j)})} \tag{3.14}$$

The intensity based weighting matrices were then re-scaled according to equation 3.15, where b is a scalar value between (0 , 1) which will be referred to as the 'control variable' henceforth. Therefore the values in the weighting matrix $\nabla\mathbf{I}_{(i,j)}$ will vary in the interval of (b , 1).

$$\nabla\mathbf{I}_{(i,j)} \;=\; \nabla\mathbf{I}_{(i,j)} * (1 - b) \;+\; b \tag{3.15}$$

The purpose of the control variable b is to control the influence from the intensity variation on to the standard Laplacian calculation. A lower value for b will increase the influence of the intensity information and vice versa. Therefore, a value of 1 for the control variable will calculate the standard Laplacian vector field with no influence from the intensity values. This value has to be set based on how prominent and consistent the intensity patterns are in a given image. Practical range of values lied between the limited range of (0.7 , 1) for this experiment. Control variable values less than 0.7 was observed to deviate the vector field in order to follow pale paths (along the centerline of the chromosome) created by the presence of sister chromatid separation. Therefore, empirically the control variable b was set to 0.9 for all the experiments in this research due to the limited banding information present in DAPI and Giemsa stained cell images as opposed to staining methods such as G-banding used in Karyotype analysis.

Once these sets of intensity weighting factor matrices were calculated, those values were directly used to change the way the Laplacian static field was calculated at each iteration. Therefore the kernel given by table 3.1 was replaced by the kernel in table 3.2, which was then defined for each (x , y) coordinate location in the image. It is important to notice that the sum of all the elements in the modified kernel does not equal to zero in this implementation. Therefore, this yielded a static vector field generation process that included both non-uniform and local shape features depending on the intensity variation in the region and the control variable b which controlled the amount of biasing. When using the standard Laplacian kernel, every pixel influenced the 8 connected neighbors uniformly. However in the proposed method, each pixel affected the neighboring pixels based on the intensity similarity or difference between them. Its also important to realize that these weight matrices are static in nature and do not change with each iteration. Therefore the Laplacian kernels needed to be calculated only once per image window.

Table 3.2: The kernel that integrates intensity information into the Laplacian calculation for location (x, y) in a $3x3$ local neighborhood.

$-\frac{\nabla\mathbf{I}_{(-1,1)}(x,y)}{8}$	$-\frac{\nabla\mathbf{I}_{(0,1)}(x,y)}{8}$	$-\frac{\nabla\mathbf{I}_{(1,1)}(x,y)}{8}$
$-\frac{\nabla\mathbf{I}_{(-1,0)}(x,y)}{8}$	$+1$	$-\frac{\nabla\mathbf{I}_{(1,0)}(x,y)}{8}$
$-\frac{\nabla\mathbf{I}_{(-1,-1)}(x,y)}{8}$	$-\frac{\nabla\mathbf{I}_{(0,-1)}(x,y)}{8}$	$-\frac{\nabla\mathbf{I}_{(1,-1)}(x,y)}{8}$

Figure 3.17 depicts the difference between the standard Laplacian kernel and (one instance of) the proposed intensity based Laplacian kernel. The instance of the proposed kernel (figure 3.17(d)) clearly depicts the biasing of the Laplacian field in directions where similar intensities are present.

Once the steady state was achieved, the gradients at each pixel location (Φ) were calculated along the two major axes (x and y) using neighborhood pixel values as given below where $B(x, y)$ was the steady state Laplacian image,

$$\begin{aligned}\frac{\Phi(x,y)}{\Delta x} &= \frac{(B(x+\Delta x,y)-B(x-\Delta x,y))}{2}\\ \frac{\Phi(x,y)}{\Delta y} &= \frac{(B(x,y+\Delta y)-B(x,y-\Delta y))}{2}\end{aligned} \tag{3.16}$$

Then each of these gradient components were normalized and stored in matrices N_x and N_y using the magnitude of the vector at each pixel. The matrices N_x and N_y contained the intensity biased Laplacian static field vector components for x and y axis directions.

Once the proposed intensity integrated Laplacian static field was derived, the corresponding contour points and the distance between them was calculated. The same thickness measures was obtained by using starting points from either contour segments or even the centerline points of the chromosome. The Euler's method was used for the above task. This is a simple and yet effective way of traversing through a vector field as given by equation 3.17, based on the local vector field direction and magnitude. For implementation, the direction of traverse had to be adjusted at times (by flipping the polarity of the vector field when necessary) in order to assure that the thickness was measured within the chromosome body. The Euler's method is an incremental method which utilizes the gradient information as given by equation 3.17 below,

$$\begin{aligned}x_n &= x + \Delta x\\ y_n &= y + y'\Delta x\end{aligned} \tag{3.17}$$

Therefore the gradient values at each pixel location (Ψ) was calculated next using each tangent vector components (N_x and N_y) as given by equation 3.18.

$$\Psi(x,y) = y'(x,y) = N_y(x,y)/N_x(x,y) \tag{3.18}$$

The Euler's method was used for thickness measuring of metaphase chromosomes using the following steps of operations.

1. Based on the local direction of the vector field gradient ($\Psi(x,y)$), select the direction (axis) for incrementing.

2. Apply Euler's equation and calculate the new pixel location along the direction of the vector field.

3. Calculate the Euclidean distance between the new and current location and accumulate with the current total distance.

4. If the new location is within the object of interest, start from the first step onwards. Once the calculated location placed outside the object, the algorithm will move to the next contour point.

The collection of these accumulated Euclidean length values is considered as the thickness/width profile of that object. Figure 3.18 depicts the steps of tracing the thickness at one contour location of the chromosome. This algorithm uniformly samples the thickness of the chromosome along the longitudinal axis of symmetry while mitigating the effects from noisy boundary conditions and chromosome bends.

3.6 Candidate point generation & metaphase centromere detection

Accurate detection of the centromere location in human metaphase chromosomes is a critical step in many clinical diagnosis processes. This location is visually characterized by a constriction in the width of the chromosome body. Therefore, an accurate width profile of a chromosome can be directly utilized to detect the centromere location. The majority of the existing methods resort to detecting the global minima of the measured feature profile in order to locate the chromosome centromere. However, the telomere regions have the possibility of having the smallest width value due to the general anatomy of the chromosome. Some of the methods in literature including the approach discussed in section 3.3.1 resorted to pruning the centerline or effectively the ends of the width profile in order to exclude the telomere region from consideration. However, this step has the potential to remove the ability to detect acrocentric chromosome centromere locations which are located towards the telomere regions. Hence some methods in the literature resorted to excluding selected classes of chromosomes in order to improve the accuracy of the published methods. These exclusions varied from excluding all acrocentric chromosomes [61] to including only metacentric chromosomes [62]. However, for a given algorithm to be used effectively for diagnostic purposes, it needs to be capable of handling all groups and morphologies of chromosomes.

Bends in chromosomes along with artifacts such as sister chromatid separation can cause the simple approach of detecting the centromere location at the global minima of the feature profile erroneous. A handful of methods have considered more than one location in the feature profile as a candidate for the centromere location. In one such approach, four candidate points were selected based on the minima values from the width profile [34]. However, this limits the number of possible locations that could be detected as the centromere location. Specially in cases where high degree of sister chromatid separation is present, limiting the search to just few candidate can have adverse effects.

In this research, a simple criteria is proposed to select all possible local minima locations as candidates for the centromere location in a given chromosome, as given below.

The notation p will be used to refer to point(s) in general within the chapter. Let the contour C be partitioned into two segments C^1 (starting segment for tracing lines) and C^2 (see figure 3.19). The width measurement of the normalized width profile (say n number of width values) at the discrete index λ ($W(\lambda)$) is obtained using the trace line which connects the contour points C^1_λ and C^2_λ from the two contours C^1 and C^2.

Then the set of candidate points for the centromere location p^C (which stores the indices λ) were calculated where either of the two logical criterions below were satisfied,

1. where the local minima conditions of $W(\lambda - 1) < W(\lambda) < W(\lambda + 1)$ and $W(\lambda - 2) < W(\lambda) < W(\lambda + 2)$ are met for all valid locations λ of the width profile (i.e. $\lambda = 3, ..., n - 2$).

2. where the local minima conditions of $W(\lambda - 1) < W(\lambda) < W(\lambda + 1)$ is met at $\lambda = 2, \ \lambda = n - 1$ of the width profile.

The second criteria was placed to account for acrocentric chromosomes where the centromere location is located towards one extreme of the width profile. However if either of the above criterions failed to yield at least one candidate location, the global minima was selected by default as the sole candidate. This was observed to occur rarely in chromosomes which are short and stubby that the width profile was not wide enough to yield a candidate. Next, the following two sets of indices were created to correspond with each given element $p^C(\alpha)$ of p^C,

- $p^{mL}(\alpha) = W(\beta)$ where $W(\beta) > W(\gamma),\ \forall\, \gamma < p^C(\alpha)$. Here $p^{mL}(\alpha)$ stores the index of the global maxima for the portion (referred to as a regional maxima henceforth)of the width profile prior to the candidate minima index $p^C(\alpha)$.
- $p^{mR}(\alpha) = W(\beta)$ where $W(\beta) > W(\gamma),\ \forall\, \gamma > p^C(\alpha)$. Similarly $p^{mR}(\alpha)$ stores the index of the global maxima for the portion of the width profile after the candidate minima index $p^C(\alpha)$.

Once the centromere candidate points p^C and their corresponding maxima points p^{mL} and p^{mR} were calculated, the set of features F^c were calculated as given below. A set of 11 features F^c were proposed to train the third SVM classifier which was then used to calculate the best candidate for a centromere location in a given chromosome. Features F_1^c to F_3^c provide an insight on the significance of the candidate point with respect to the general width profile distribution. The normalized width profile value itself is embedded in features F_4^c and F_8^c where the latter scales the minima based on the average value of the width profile. Features F_5^c and F_6^c capture the contour curvature values that are intrinsic to the constriction at the centromere location. Features F_7^c,F_9^c and F_{10}^c include distance measures which indicate the positioning of the candidate point with respect to the chromosome as well as to the width profile shape. Finally the feature F_{11}^c records the staining method used in the cell preparation. This gives the classifier some crucial information that is then used to accommodate for specific shape features introduced through the cell preparation technique used.

Let i be a candidate member number in the centromere candidate pool. Also let $d(1, i)$ be the Euclidean distance along the midpoints of the width profile trace lines (centerline) from a telomere to the candidate point and L be the total length of the chromosome. Then the set of features F^c are stated as below where $\|.\|$ yields the absolute value,

1. $F_1^c = \Big\| W(p^C(i)) - W(p^{mL}(i)) \Big\|$. This feature calculates the absolute width profile difference between the candidate and the regional maxima prior to the candidate point on the width profile.

2. $F_2^c = \left\| W(p^C(i)) - W(p^{mR}(i)) \right\|$. This feature calculates the absolute width profile difference between the candidate and the regional maxima beyond the candidate point on the width profile.

3. $F_3^c = F_1^c + F_2^c$ which calculates the combined width profile difference created by the candidate point.

4. $F_4^c = W(p^C(i))$. This captures the value of the width profile ($0 \leq F_4^c \leq 1$) at the candidate point location.

5. F_5^c is the local curvature value at the contour point C_λ^1 which corresponds to the current centromere candidate location (where $\lambda = p^C(i)$)

6. F_6^c is the local curvature value at the contour point C_λ^2 which corresponds to the current centromere candidate location (where $\lambda = p^C(i)$)

7. $F_7^c = min\left(d(1,i), L - d(1,i)\right)/L$. Gives a measure where the candidate is located with respect to the chromosome as a fractional measure ($0 \leq F_7^c \leq 0.5$)

8. $F_8^c = W(p^C(i))/\bar{W}$, where $\bar{W}$ is the average of the width profile of the chromosome. This includes the significance of the candidate point minima with respect to the average width of the given chromosome.

9. $F_9^c = d(p^{mL}(i), p^C(i))/L$. This gives the distance between the candidate point location and the regional maxima value prior to the candidate point, normalized by the total length of the chromosome.

10. $F_{10}^c = d(p^C(i), p^{mR}(i))/L$. This gives the distance between the candidate point location and the regional maxima value beyond the candidate point, normalized by the total length of the chromosome.

11. F_{11}^c is a boolean feature used to indicate the staining process used during cell preparation. A logical '0' would indicate the use of DAPI staining while '1' would indicate a Giemsa stained cell.

The detection of the centromere location assumes that each chromosome at least contains one centromere location within the chromosome. This is a reasonable assumption since the centromere region is an integral part of the anatomy which is normally retained in cell division with the exception of cases with excessive radiation exposure. This assumption transforms the detection problem into a ranking problem in which we pick the best out of a pool of candidates. Therefore the same approach that was utilized for the contour partitioning algorithm (section 3.4)was adopted here in which the distance from the separating hyperplane ρ (the geometric margin) was used as a measure of goodness of fit for a given candidate. This metric reduced the multidimensional feature space in to a single dimension metric which inherently reduced the complexity in ranking the candidate locations. Since the large margin binary classifier (SVM) oriented the separating hyperplane in the feature space, the 1D distance metric was directly related to how well a given candidate fits into the general characteristics of a given class label. A detailed principal component analysis (PCA) of the centromere detection features F^c is provided in the following section followed by an introduction to the candidate based centromere confidence metric.

3.6.1 Principal component analysis (PCA) of features

PCA is a technique used for mapping a set of correlated features into principal components using orthogonal basis functions where each component captures the variance of the data set in ascending order. Therefore PCA is commonly used in literature to analyze and reduce the number of features in many machine learning problems. PCA was performed in this research for the feature set used for centromere detection in order to obtain an insight into the feature set through their contribution to the overall variance. Table 3.3 provides the percentage figures of contribution to variations (in descending order) from the set of features. The candidate width profile value (F^c_4) and the width profile value scaled by the average width value (F^c_8) accounts for almost 50% of the variation. The first 8 features in table 3.3 accounts for over 95% of the total variance while leaving out the curvature features and the image staining method features. However, when reduced to these 8 features, both the classification accuracy as well as the detection accuracy deteriorated noticeably. Here it was surmised that the inclusion of the maximum variance through PCA does not guarantee better separation of data in the feature space. Therefore, the original 11 features were retained in the experiment.

Table 3.3: The percentage contribution of each feature to the variance of the whole data set in the descending order.

Number	Feature number	Percentage of variance %
01	F_4^c	26.67
02	F_8^c	23.26
03	F_1^c	14.16
04	F_2^c	11.68
05	F_7^c	7.56
06	F_3^c	6.09
07	F_9^c	4.17
08	F_{10}^c	3.03
09-11	$F_5^c + F_{11}^c + F_6^c$	3.38

3.6.2 Candidate based centromere confidence (CBCC)

Centromere localization is an essential stage in many cytogenetical diagnosis processes. Although the accuracy measures are a good basis for establishing performance of a machine learning application, it does not yield any useful information regarding the accuracy of a given classification or localization. A measure of confidence in a particular localization is a useful tool in diagnosis processes. Therefore an intuitive and effective measure termed the Candidate Based Centromere Confidence (CBCC) for representing the confidence was proposed in the detection of a centromere location, which can be obtained using the solution space derived through the classifier and the distance metric ρ.

For a given set of candidate points of a chromosome p^C, the goodness of fit (GF) of the optimal candidate point ($\grave{\rho}$) is obtained by calculating $\left\|\frac{(\grave{\rho}-\bar{\rho})}{2}\right\|$ which is the average distance of all the remaining candidate points. The expected detection would have the optimal candidate as well as the other candidates as support vectors for the classifier on either side of the separating hyperplane (see figure 3.20). The average of the rejected candidates was used instead of the second best candidate for gauging the goodness of fit due to multiple reasons. First, it was possible to have multiple candidates within a given centromere region. Also there are dicentric chromosomes that possess two centromere locations which could adversely affect the CBCC value if the second best candidate was used as a benchmark. Here, the optimal candidate distance ($\grave{\rho}$) is ≈ 1 while the average of the remaining candidate distances ($\bar{\rho}$) be ≈ -1. The GF value was truncated at unity since exceeding this value does not add any new information to the measure.

Table 3.4 depicts CBCC values for accurately detected chromosomes as opposed to inaccurately detected chromosomes. It also includes a third category termed ”All nonviable candidates chromosomes” (a subset of inaccurate centromere detection) where none of the candidates of a given chromosome were marked as capturing the true centromere of the chromosome. These were mainly caused by segmentation of acrocentric chromosomes where the lighter intensity of the satellites were segmented out and also by extreme sister chromatid separation. Values in table 3.4 shows a clear correlation between the CBCC values and the accuracy of the centromere localization outcome and therefore representing the confidence in the detection.

Table 3.4: The mean and the standard deviation of the CBCC values in cases with accurate centromere detection as well as inaccurate centromere detection. The table also includes cases where non of the candidates were not found to be viable candidates (a subset of inaccurate centromere detection) for the centromere location.

Category	Mean (μ)	Std. Dev (σ)
Accurate detection	0.7861	0.3000
Inaccurate detection	0.3799	0.3293
All nonviable candidates	0.2696	0.2457

3.6.2.1 Customizing the confidence value

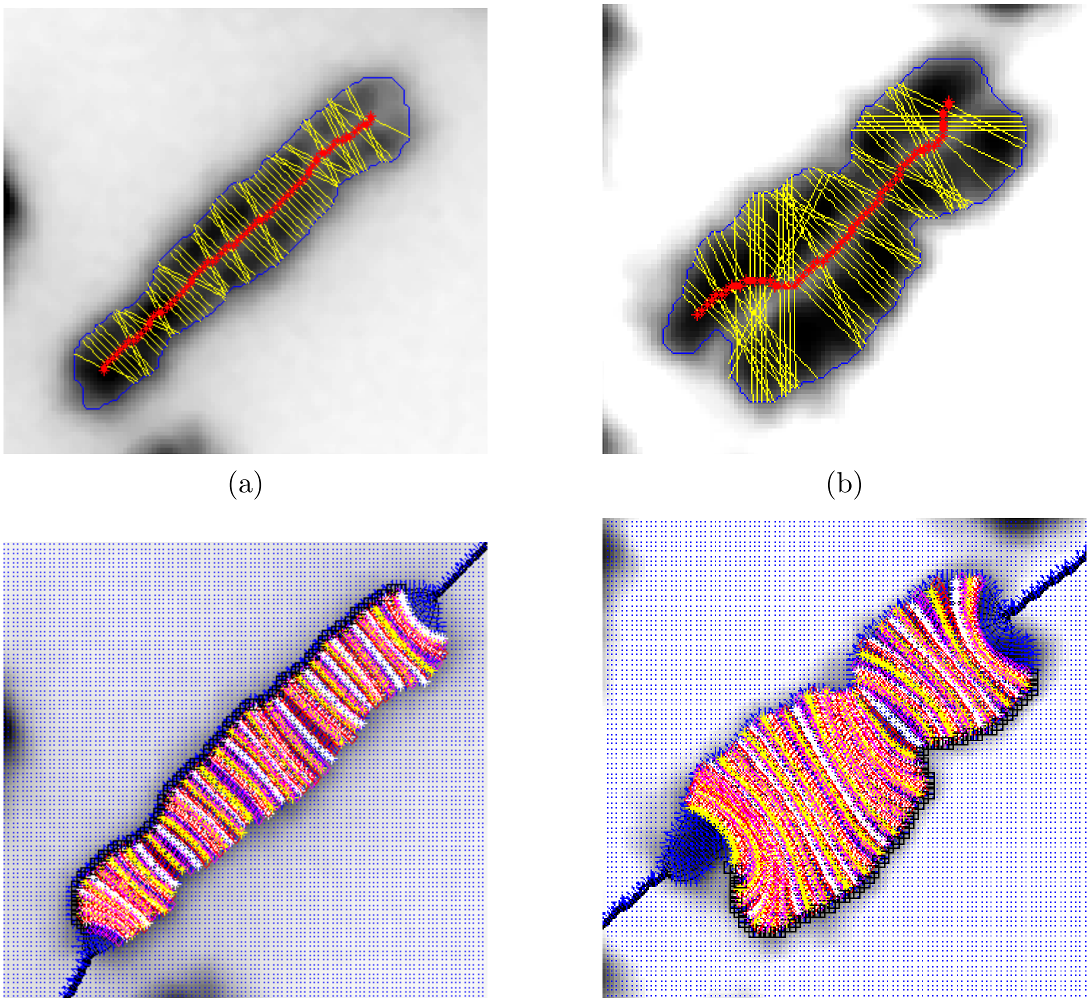
In clinical diagnosis applications and in many other candidate based approaches, a percentile value of the detection confidence could convey information better regarding a detection result in a more comprehendible form. However it is important to note that a given certain percentile value (say 90%) would imply two different levels of significance depending on the application. For an example in critical diagnosis problems, a 90% value should provide very high confidence in the measurement which may not be demanded by another candidate based detection problem. Therefore, this percentile measure is highly subjective and problem specific unlike the CBCC measure discussed in section 3.6.2. This also implies the requirement of a mapping framework which can be easily altered for creating any problem specific confidence measurement schema.

This is proposed to be obtained by mapping the CBCC value prior to truncation (GF) using the cumulative distribution function of a normal distribution. The mean and the standard deviation values used for this gives a control to the user to set the sensitivity of the accuracy values to match the required standards and then filter out detections below a certain threshold CBCC value. For an example, in situations where detection accuracy is highly important the user can fine tune the confidence percentile value by setting a relatively higher mean value and a lower standard deviation value. This assigns a lower value for cases with low goodness of fit and focuses the major segment of mapping to the higher GF region. The spread of this region is controlled through the standard deviation value. Figure 3.21 demonstrates the curve shape of few possibilities of cumulative distribution functions that can be used to map the accuracy values according to the requirements of the given problem. This improves the applicability of this measure to various applications not only in centromere detection, but for any application which involves in selecting an optimal candidate from a set of candidates. Furthermore, this scheme utilizes the CBCC values which are beyond 1.00 and rank these ideal cases to fit into the higher percentile value range.

Since centromere detection carries important information regarding the anatomy of the chromosome, the mean and the standard deviation values for instance were set at 0.75 and 0.5 respectively which is represented by figure 3.21(c). This penalized detections with less desirable GF while mapping most of the cases with higher GF properly to the percentile confidence values. Table 3.5 depicts these mapped percentile values (derived using figure 3.21(c)) for accurately detected chromosomes, inaccurately detected chromosomes and all nonviable candidates chromosomes similar to table 3.4. A close examination of the values in table 3.5 in comparison to those in table 3.4 for the same groups reveals the effect of normalization (mapping). A more aggressive mapping function as given by figure 3.21(d) would increase the difference between the mean values of correct and incorrect localization values while further penalizing the lower CBCC values.

Table 3.5: The mean and the standard deviation of the percentile confidence values in cases with accurate centromere detection as well as inaccurate centromere detection. The table also includes cases where none of the candidates were not found to be viable candidates (a subset of inaccurate centromere detection) for the centromere location.

Category	Mean (μ)	Std. Dev (σ)
Accurate detection	70.01%	24.34%
Inaccurate detection	29.29%	27.64%
All nonviable candidates	19.29%	14.44%

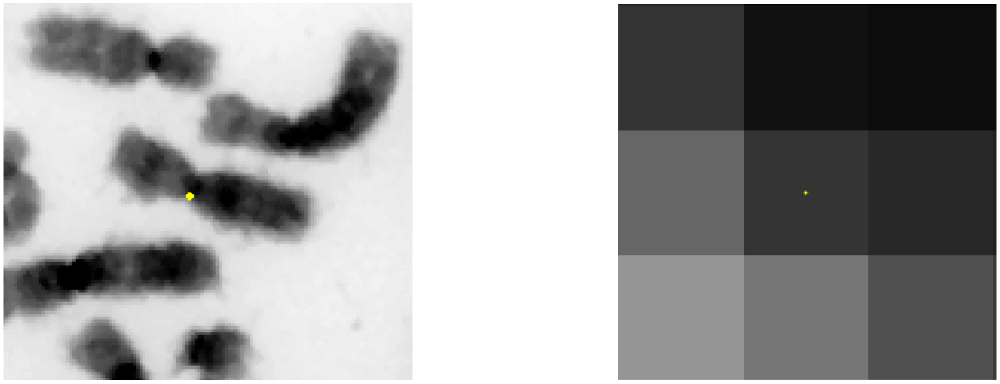

(a)

(b)

(c)

(d)

Figure 3.16: Depicts the uniform sampling of the width profile using the proposed method in figure 3.16 (c) & (d) as opposed to the trellis structure measurements through the centerline based approaches given by figure 3.16 (a) & (b).



(a) (b)

$$\begin{matrix} -0.1250 & -0.1250 & -0.1250 \\ -0.1250 & 1.0000 & -0.1250 \\ -0.1250 & -0.1250 & -0.1250 \end{matrix}$$

(c)

$$\begin{matrix} -0.1271 & -0.1253 & -0.1250 \\ -0.1246 & 1.0000 & -0.1265 \\ -0.1221 & -0.1238 & -0.1257 \end{matrix}$$

(d)

Figure 3.17: Demonstrates the difference between the kernel of the proposed method in comparison to the standard Laplacian kernel. Figure 3.17(b) is an enlarged view of the 3x3 neighborhood of the pixel location marked by yellow on figure 3.17(a). Figure 3.17(c) & (d) represents the standard Laplacian and intensity biased Laplacian kernels calculated for the neighborhood of interest.

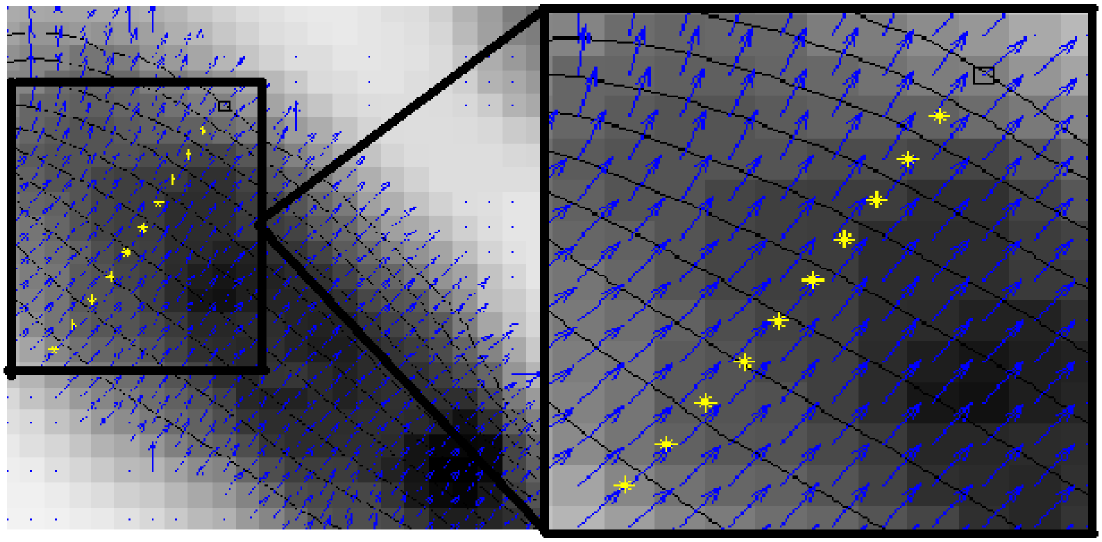


Figure 3.18: The steps of tracing the thickness (yellow stars) at one contour location of the chromosome where the arrows indicate the Laplacian vector field. The black square indicates the end point on the contour of the object. The final thickness value is calculated by getting the sum of all the lengths of these small steps.

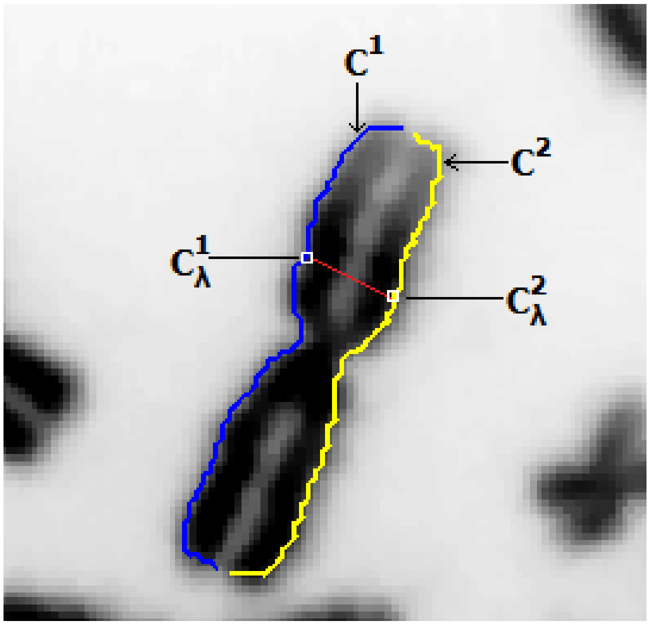


Figure 3.19: An example where the contour C is split into two approximately symmetric segments C^1 and C^2. The red width trace line connects the points C^1_λ and C^2_λ of the two contour segments.

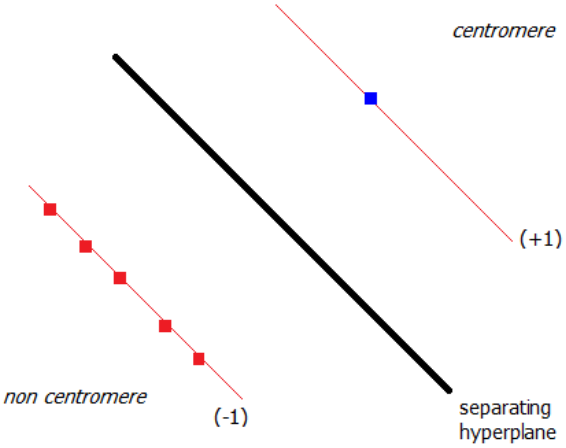


Figure 3.20: Shows the expected scenario for candidate based centromere detection where the blue square represents the optimal candidate while the other five candidates are given by the red squares in the feature space.

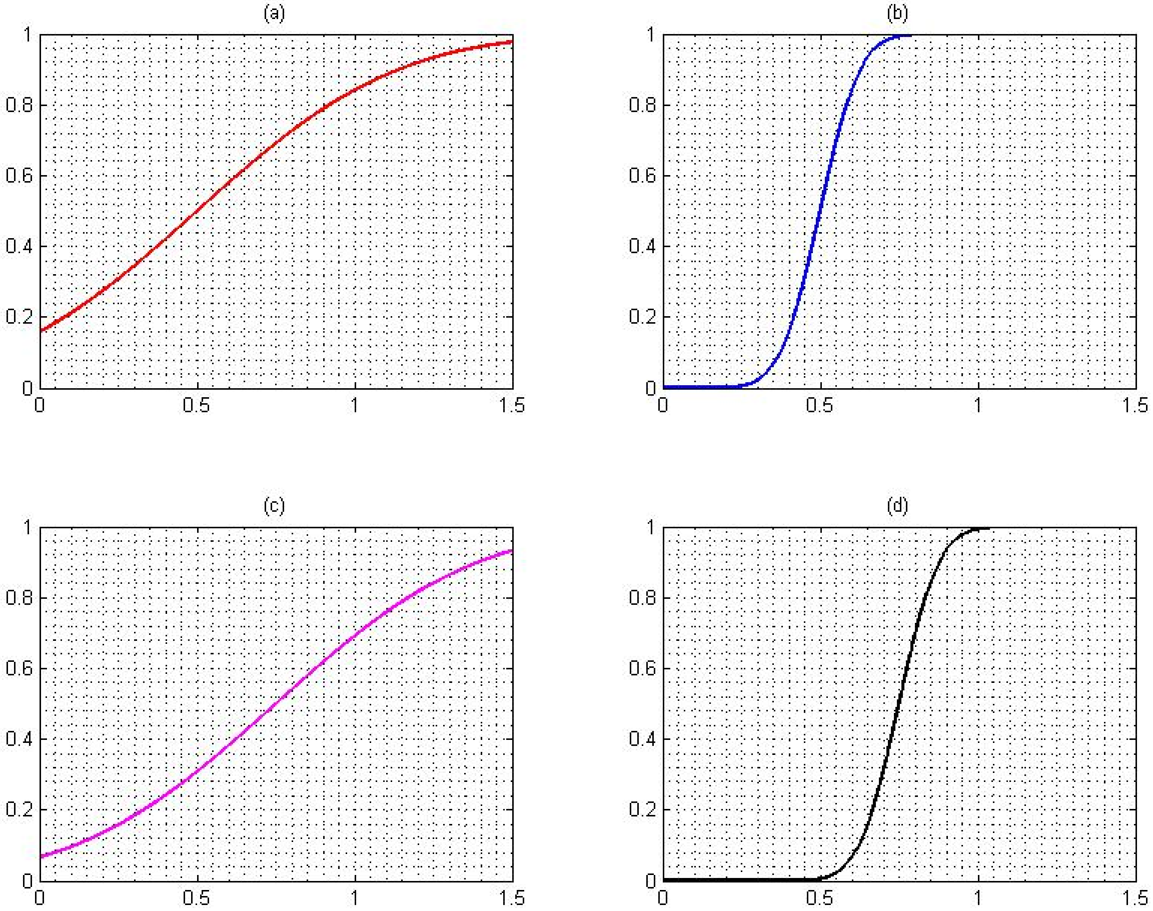


Figure 3.21: A collection of possible cumulative distributive functions (of Gaussian distributions) where the (μ, σ) values are set at (0.5, 0.5) for figure 3.21(a) , at (0.5, 0.1) for figure 3.21(b), at (0.75, 0.5) for figure 3.21(b) and finally at (0.75, 0.1) for figure 3.21(d).

Chapter 4 Results

Centromere localization is an essential step in many chromosome analysis algorithms which can be used to derive information such as the Centromere Index (CI), chromosome group and chromosome number. Therefore the centromere detection accuracy was used to measure the performance of the proposed algorithm. Testing was carried out using two different test schemas designed to test different aspects of the proposed algorithm as listed below.

- **Preliminary testing :** The main objective of this stage of testing was to analyze the performance of the proposed method in accurate sampling of the width profile of the chromosome. The performance of the proposed method was tested against a centerline based method and an in depth statistical analysis was conducted to establish the statistical significance of the observed results.
- **Candidate based centromere detection :** This stage of testing analyzed the candidate based centromere detection accuracy of the proposed method. Testing was done based on a sizable and diverse data set of chromosome which contained all groups of chromosomes with different staining methods.

4.1 Preliminary testing

The main objective of this test was to establish the accuracy in sampling of the width profile of the proposed method using the centromere detection accuracy. The centromere and centerline detection method described in section 3.3.1 was used for comparing the results of the proposed method. The centerline based method was selected since it can handle any chromosome morphology without yielding spurious branches in the centerline and also attempts to find the width profile similar in principle to most existing methods [1],[41]. The results of the proposed method was tested for statistical significance against those obtained using the centerline approach.

Since the centerline based method used a pruned centerline for width profile calculation, the width profile of the proposed method was also pruned by the same extent at each end. Furthermore, both methods were set up to detect the centromere location by calculating the global minima of the width profile. The data set for this stage consisted of 226 human lymphocyte chromosomes from 12 chromosome cells which yielded an average of 18.8 chromosomes per cell image where only non overlapping and touching chromosomes were selected. Due to the limitations of the centerline based method, a majority of the data set consisted of metacentric and submetacentric chromosomes with skeletons longer than 35 pixels were included in the analysis. Cell images contained chromosomes with different staining methods and artifacts. Table 4.1 provides the breakdown of these cell images based on the staining method as well as the presence of sister chromatid separation (judged visually).

Table 4.1: Breakdown of chromosome cell images and chromosomes based on the staining method and the sister chromatid separation (SC Sep.)

Abbr.	**Label**	**Images**	**Chromosomes**
D-NSC	DAPI-No SC Sep.	4	72
D-WSC	DAPI-With SC Sep.	3	59
G-WSC	Giemsa-With SC Sep.	5	95
	Total	12	226

A proper mechanism for collecting ground truth was essential in order to quantitatively measure the accuracy of centromere detection using both centerline and the Laplacian based (proposed) methods. For this experiment, the centromere location manually recorded by the author was used as the 'gold standard' in the analysis. Due to limitations of resources, the intra-observer variability of ground truth was not analyzed in the current stage of the research. Here, the ground truth was collected in the form of a line drawn across the centromere region. Then, the perpendicular distance in pixels from the centromere location given by the algorithm (midpoint of the scan line with the minimum width) to the user drawn line segment was denoted as the error of detection. In this study, the interest was placed in errors in the vicinity of the center of the chromosome rather than the orientation of the scan line. Therefore the experiment was set up in a way that any displacement of the detected centromere location along the drawn ground truth centromere line would not influence the accuracy of detection. Furthermore, pixel error values were not normalized since the centromere structure mostly remains fixed despite chromosome morphology or chromosome number. These error values will be denoted by E_L and E_C for the error of the Laplacian based proposed method centromere and the centerline method result respectively.

The Laplacian based proposed algorithm performed well on chromosomes despite the staining method and the shape of the chromosome which was verified with the statistical analysis performed in section 4.1.1. However, the algorithm failed in the presence of high sister chromatid separation in the binary segmentation of the chromosome. The DCE algorithm in contour partitioning selected a high curvature point within the telomeric region and thus yielded the correction for sister chromatid separation ineffective. Figure 4.1 shows some of the sample results for multiple staining methods used commonly in cytogenetic studies. Figure 4.1 (f) depicts an instance where the correction for sister chromatid separation had failed to yield the expected result. In figure 4.1 (e) a case is presented where the centerline based method had failed to yield an accurate centromere location. This was caused by a noisy centerline which missed the actual width constriction at the centromere location. However the proposed method yielded better results due to uniform sampling despite the object boundary noise (see figure 4.2 (d)).

4.1.1 Statistical analysis

Once error measurements were calculated, statistical analysis techniques were used to reject or accept the null hypothesis which states that both the proposed Laplacian error measurements (E_L) and the centerline method error measurements (E_C) were from the same population. Therefore the failure to reject the null hypothesis concluded a statistically insignificant improvement from the proposed Laplacian based method. However, the following conditions needed to be fulfilled before applying any parametric statistical analysis test.

1. The samples of both error distributions should contain normal distributions.

2. Both error distributions have equal variance.

3. Chromosomes used for testing are selected through a complete random process.

4. Both groups contain equal number of chromosomes (same sample size).

The normality of the two error distributions were first examined using the descriptive values given in table 4.3. The proposed algorithm yielded a smaller error

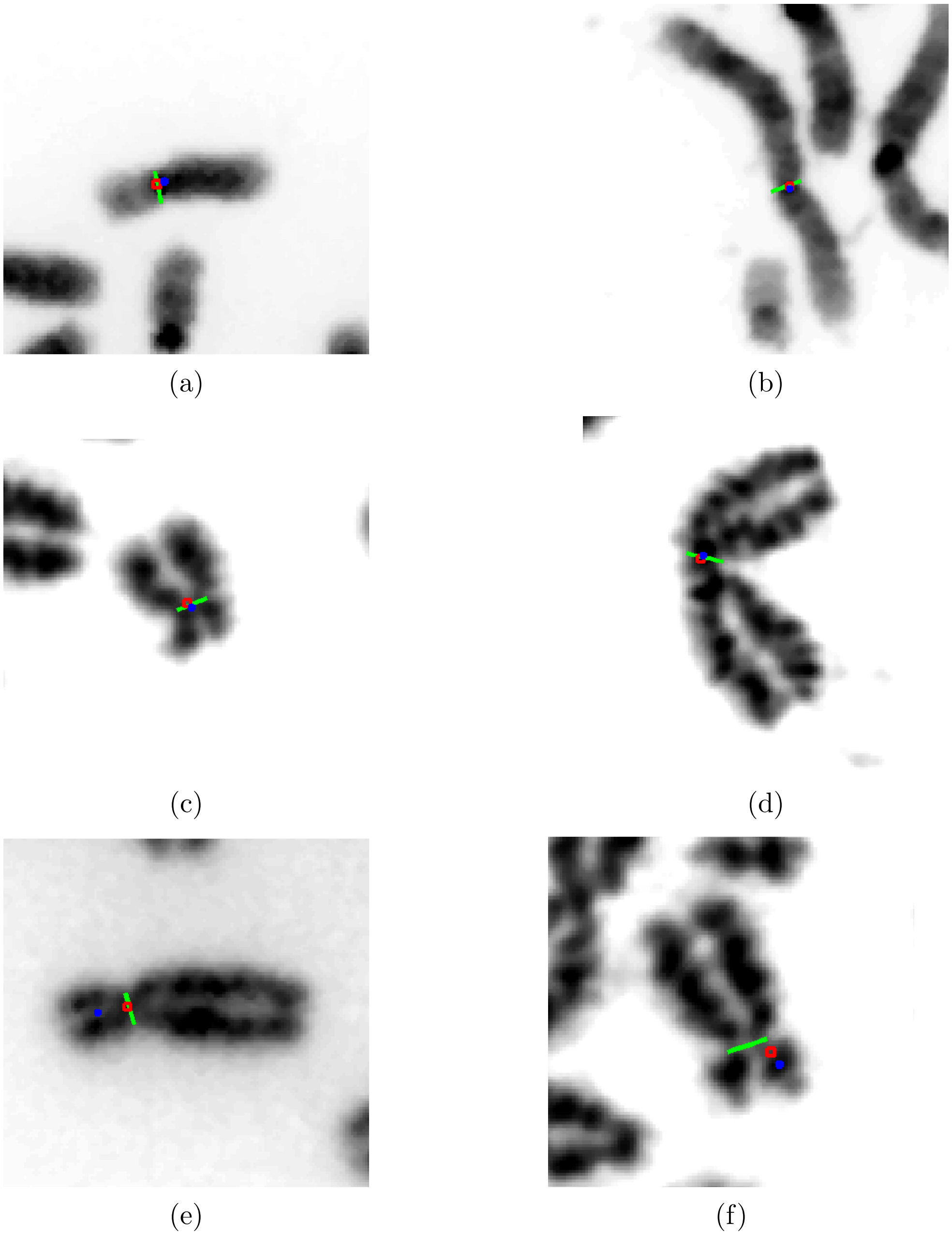
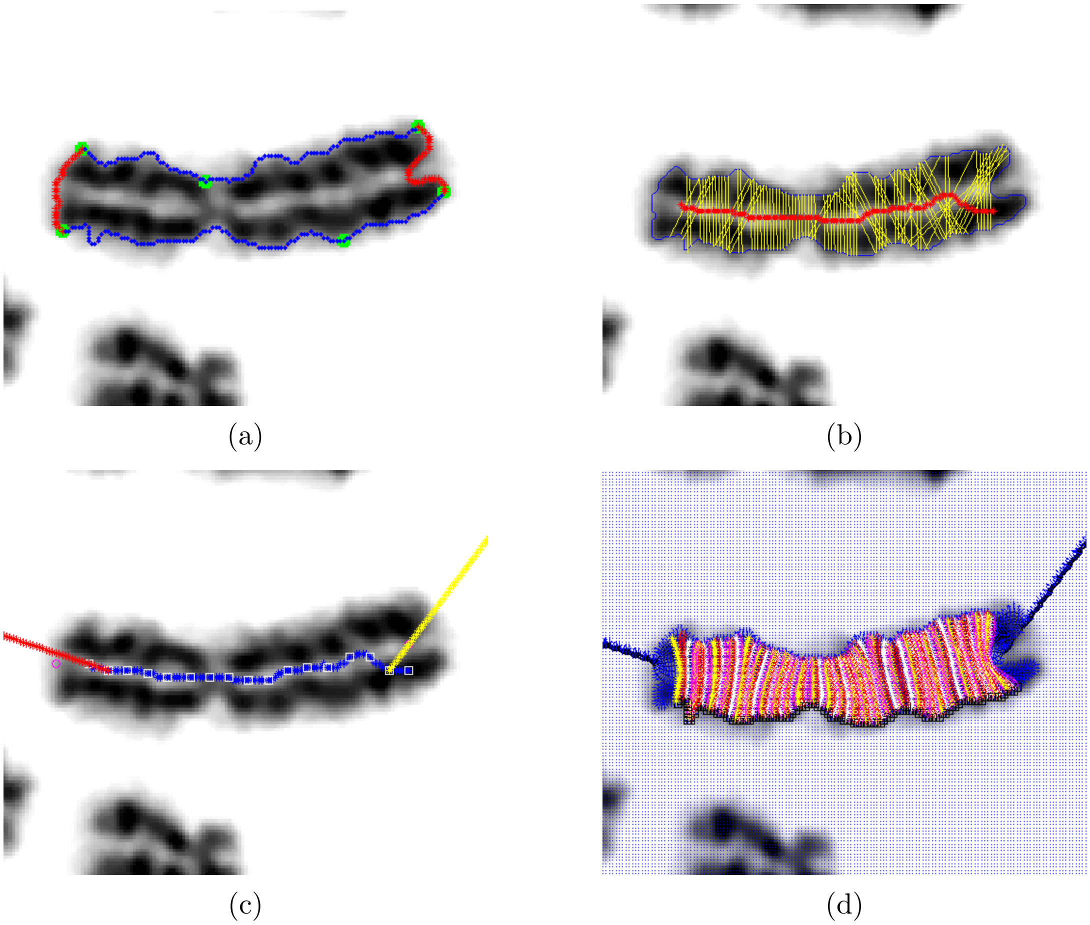


Figure 4.1: Demonstrates some sample results of the algorithm where the detected centromere location is depicted in red color circle against the centerline based approach [1] in blue color star while the ground truth centromere line is depicted in white. Figure 4.1 (a)&(b) are results of DAPI (4',6-Diamidino-2-Phenylindole) stained chromosomes while figure 4.1 (c)-(f) are results of Giemsa stained chromosomes. Figure 4.1(e) is an instance in which the proposed algorithm has outperformed the state of the centerline method significantly while figure 4.1 (f) is an instance in which the proposed algorithm has failed to yield the accurate centromere location due to high degree of sister chromatid separation.



(a)

(b)

(c)

(d)

Figure 4.2: Depicts an example of the correction for the sister chromatid separation artifact and the impact of that correction on the width profile measurement in the proposed method. The right hand side telomere region was corrected for sister chromatid separation in figure 4.2 (c) where the resulting uniform sampling of the width profile from the Laplacian based method is given in figure 4.2 (d) as opposed to the centerline based method in figure 4.2 (b).

mean value with a smaller standard error of mean while having a relatively higher skewness and kurtosis values obtained in comparison to the centerline method. However the high kurtosis and skewness values in table 4.3 suggested that both distributions deviated from normality. Next, the Kolmogorov-Smirnova test was used to statistically test the normality of these distributions. The results of this test are given in table 4.2. According to this test result, the null hypothesis that these distributions were normal ($p < 0.05$) was rejected. Although this deviation from normality is not statistically preferred, the high kurtosis and skewness values along with the relatively high Kolmogorov-Smirnova statistic value ($Z = 0.280$) provided evidence towards the conclusion that the proposed method yielded a better grouped distribution towards a lower mean error value when compared to the centerline method.

Table 4.2: The Kolmogorov-Smirnova normality test results for the data set.

	Kolmogorov-Smirnova		
	Statistic	df	Sig.
E_L	.280	226	.000
E_C	.258	226	.000

Table 4.3: Descriptive values for the detection error data set when analyzed with proposed Laplacian based method (E_L) and Centerline based method (E_C) [1].

	N	Mean		Kurt-	Skew-
		Stat.	Std. Error	-sis	-ness
E_L	226	4.0243	.4535	17.859	3.839
E_C	226	8.7819	.7749	2.657	1.834

Levene's test for equal variance was then used in order to test for the null hypothesis of equal variance in the two distributions. The results of the test given in table 4.4 indicated that the null hypothesis of equal variance could be rejected ($p < 0.05$).

Table 4.4: Levene's test for testing equal variance of detection error within image groups (given in table 4.1) for each algorithm.

	Levene Statistic	df1	df2	Sig.
E_L	9.763	2	223	.000
E_C	23.362	2	223	.000

Earlier mentioned 3rd and 4th conditions had to be ensured during the design stage of the experiment. The 4th condition about having the same sample size was met since both algorithms were tested against the same number of chromosomes. Although these chromosome were selected without any bias, this test could not be considered as a random process since the same set of images were presented for both algorithms. Therefore these experiments violated the third condition and fell under the category of 'repeated measurement' analysis. Therefore, a 't statistic' cannot be utilized to obtain the significance of the results.

Since the experimental setup as well as the error distributions violates some of the assumptions of parametric statistical analysis, nonparametric methods were selected to test for the statistical significance of the proposed method. This was obtained through the 'Wilcoxon Signed Rank Test' which could be directly applied to repeated measurements without the normality constraint. The results of the test are given below in table 4.5.

Table 4.5: The Wilcoxon signed test rank analysis results.

		N	Mean Rank	Sum of Ranks
E_C - E_L	(-) Ranks	80	89.74	7179.00
	(+) Ranks	146	126.52	18472.00
	Ties	0		
	Total	226		

The data in table 4.5 analyzes cases based on their signs after comparing each corresponding pair. Therefore it was observed that the sum of positive ranks were significantly higher than that for negative ranks, which corresponded to the cases when the Laplacian based proposed method gave lower error values compared to the centerline based approach.

Table 4.6: The Wilcoxon signed test significance analysis results.

	E_C - E_L
Z	-5.738
Asymp. Sig.(2-tailed)	.000

The significance analysis of the Wilcoxon signed rank test given in table 4.6 demonstrates that the error values yielded by the proposed Laplacian based method are statistically 'rare'. Given this observation, it is reasonable to assume that the proposed method error measurements E_L and the centerline based error measurements were derived from two different populations. Therefore the null hypothesis that both E_L and E_C are from the same population (equal mean values) was rejected ($p < 0.05$), concluding that the proposed method elicit statistically significant improvement in centromere localization ($p < 0.05$) compared to the centerline based method.

Since the proposed method results are significant (rare), the variability of error measurements between the defined three labels given in table 4.1 for the proposed method was investigated. This was an essential step for judging the performance of the proposed Laplacian based method on different groups of images. The descriptives for each of these image groups for the proposed method results are given in table 4.7.

Table 4.7: Descriptive values for the detection error for the proposed Laplacian method (E_L) and the centerline based method (E_C) with respect to the image groups stated in table 4.1.

Method - Group		N	Mean	Std. Dev.	Std. Error
Error Laplacian	1.00	72	2.4835	3.00623	.35429
	2.00	59	3.9585	5.93717	.77295
	3.00	95	5.2328	8.91079	.91423
Error Centerline	1.00	72	3.3006	5.44868	.64213
	2.00	59	9.2070	11.38440	1.48212
	3.00	95	12.6721	13.56750	1.39200

A one-way Analysis of Variance (ANOVA) was performed and the results are given in table 4.8. ANOVA provides an insight to whether the variance between groups is larger than that would be expected through chance by comparing to the variance within groups. The results of this test depicted that there were statistically significant variations between the groups in both methods. However the lower 'F' statistic value along with the smaller margin of significance suggested that the proposed method performed more consistently than the centerline based method within all groups of data.

The results given by the one-way ANOVA test (see table 4.8) failed to reject the null hypothesis. This warranted the need of a post-hoc test which analyzes the variability between each image groups separately. However a post-hoc analysis method had to be selected carefully since the null hypothesis of equal variance was rejected

Table 4.8: ANOVA test for significance within image groups.

		df	Mean Square	F	Sig.
E_L	Between	2	154.963	3.405	.035
	Within	223	45.516		
	Total	225			
E_C	Between	2	1805.780	14.954	.000
	Within	223	120.754		
	Total	225			

based on the Levene's test results (see table 4.4). Therefore, the Games-Howell post-hoc test was chosen for this purpose since it does not rely on homogeneity of variance assumption (refer table 4.9).

Table 4.9: Games-Howell post-hoc test results for analyzing significance of variance between image groups where the groups Group 1,2,3 were D-NSC (DAPI without SC Sep.),D-WSC (DAPI with SC Sep.) and G-WSC (Giemsa with SC Sep.) respectively.

Dependent Variable	Group (I)	Group (J)	Mean Diff	Std. Err	Sig.
E_L	1.00	2.00	-1.4749	.8503	.199
		3.00	-2.7493	.9805	.016
	2.00	1.00	1.4749	.8503	.199
		3.00	-1.2743	1.1972	.538
	3.00	1.00	2.7493	.9805	.016
		2.00	1.2743	1.1972	.538
E_C	1.00	2.00	-5.9064	1.6152	.001
		3.00	-9.3715	1.5330	.000
	2.00	1.00	5.9064	1.6152	.001
		3.00	-3.4650	2.0333	.207
	3.00	1.00	9.3715	1.5330	.000
		2.00	3.4650	2.0333	.207

The post hoc analysis demonstrated that performance of the proposed method varies significantly only between D-NSC (DAPI without SC Sep.) & G-WSC (Giemsa stained with SC sep.). In the meantime, the results of centerline based method varied significantly between groups D-NSC (DAPI without SC Sep.) & D-WSC (DAPI with SC Sep.) as well as groups D-NSC (DAPI with SC Sep.) & G-WSC (Giemsa stained). Therefore it was deduced that the proposed method varied less in performance based on the image group type in comparison to the centerline based method. Furthermore, the difficulty in handling Giemsa stained images were also visible through both methods due to the high level of boundary noise introduced during the staining process. Boundary noise can often distort the width profile measurement values and inject error into centromere localization. The presence of sister chromatid separation was better handled through the proposed method compared to the centerline based method.

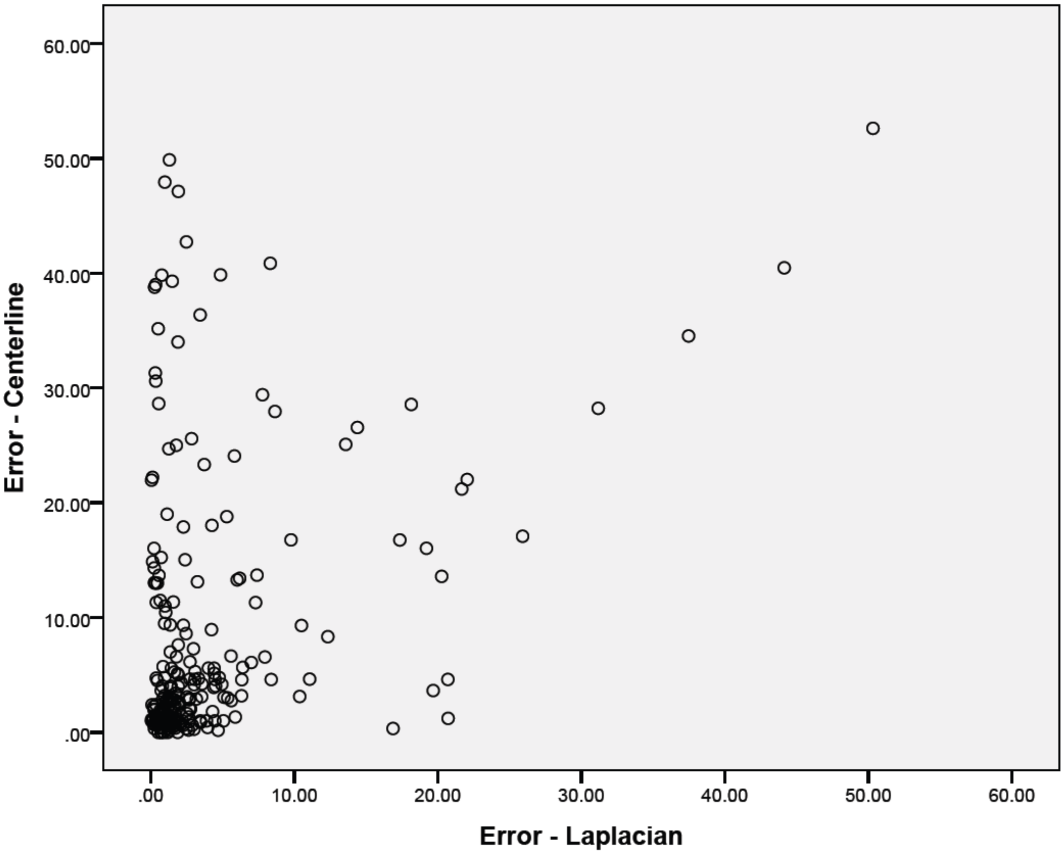


Figure 4.3: The scatter plots for demonstrating the correlation between the two detection error distributions in which the 'x' axis is the detection error of the proposed Laplacian based method (E_L) and the 'y' axis is the centerline based method (E_C).

During an analysis it is equally important to test the amount of correlation between the results obtained using the two methods. This provides insights about the similarities in results obtained from the two methods. A scatter plot is used as given by figure 4.3 for this purpose as a preliminary analysis. The outliers of the error values (a common shortcoming shared by both methods) were observed to be a result of the inaccuracy of the contour partitioning and correcting for sister chromatid separation. The same phenomenon was observed between the error values for the proposed method with and without integrating intensity information. A nonparametric correlation test using the Spearman coefficient was used to obtain a quantitative value (refer table 4.10). Therefore it was observed that there is a strong correlation between the error values obtained through both these methods. The clustering of data points in the figure 4.3 corresponds to the majority of these correlation measurements been significant.

Table 4.10: Non-parametric correlation test for significance (*Sig.*) using Spearman coefficient for detection error between the proposed algorithm and the centerline based method where $\rho_{(X,Y)}$ denotes the correlation between two random variabels.

			E_L	E_C
Spearman's rho	E_L	$\rho_{(X,Y)}$	1.000	.250
		Sig.		.000
		N	226	226
	E_C	$\rho_{(X,Y)}$	.250	1.000
		Sig.	.000	
		N	226	226

Following the statistical analysis, a preliminary study was conducted in order to explore the effects of adding the intensity information into the standard Laplacian framework. A total of 11 chromosomes were examined from the data set which had the most positive impact from the addition of intensity into the algorithm. These chromosomes on average showed an improvement of 20.9 pixels in error. However out of the 11 chromosomes, 3 were affected by the presence of high degree of sister chromatid separation. The chromosomes that were not affected by this phenomenon, showed the presence of a lighter intensity band close to the centromere location. On the other hand, only 4 chromosomes were present in the data set which considerably downgraded the result (with an average of 26.4 pixels in error). A close examination of these 4 chromosomes revealed the presence of high degree of sister chromatid separation, influencing the algorithm to detect the centromere location on one of the sister chromatids (towards the telomere region).

4.2 Candidate based method performance

It was established that the proposed algorithm outperformed centerline based approach [1] with statistical significance as demonstrated in section 4.1. In this experiment the proposed method was tested for performance on a larger data set containing 1400 chromosomes from 40 cell images (at an average of 35 chromosomes per image) containing both DAPI and Giemsa staining, with and without sister chromatid separation. Table 4.11 provides the breakdown of these cell images based on the staining method as well as the presence of sister chromatid separation (judged visually). In this experiment, the complete length of the width profile was utilized, which was a feature enabled by the use of candidates for centromere location as opposed to the global minima of the profile. Therefore, all chromosomes that were not touching or overlapping neighboring chromosomes in each cell image were included in the analysis of the experiment.

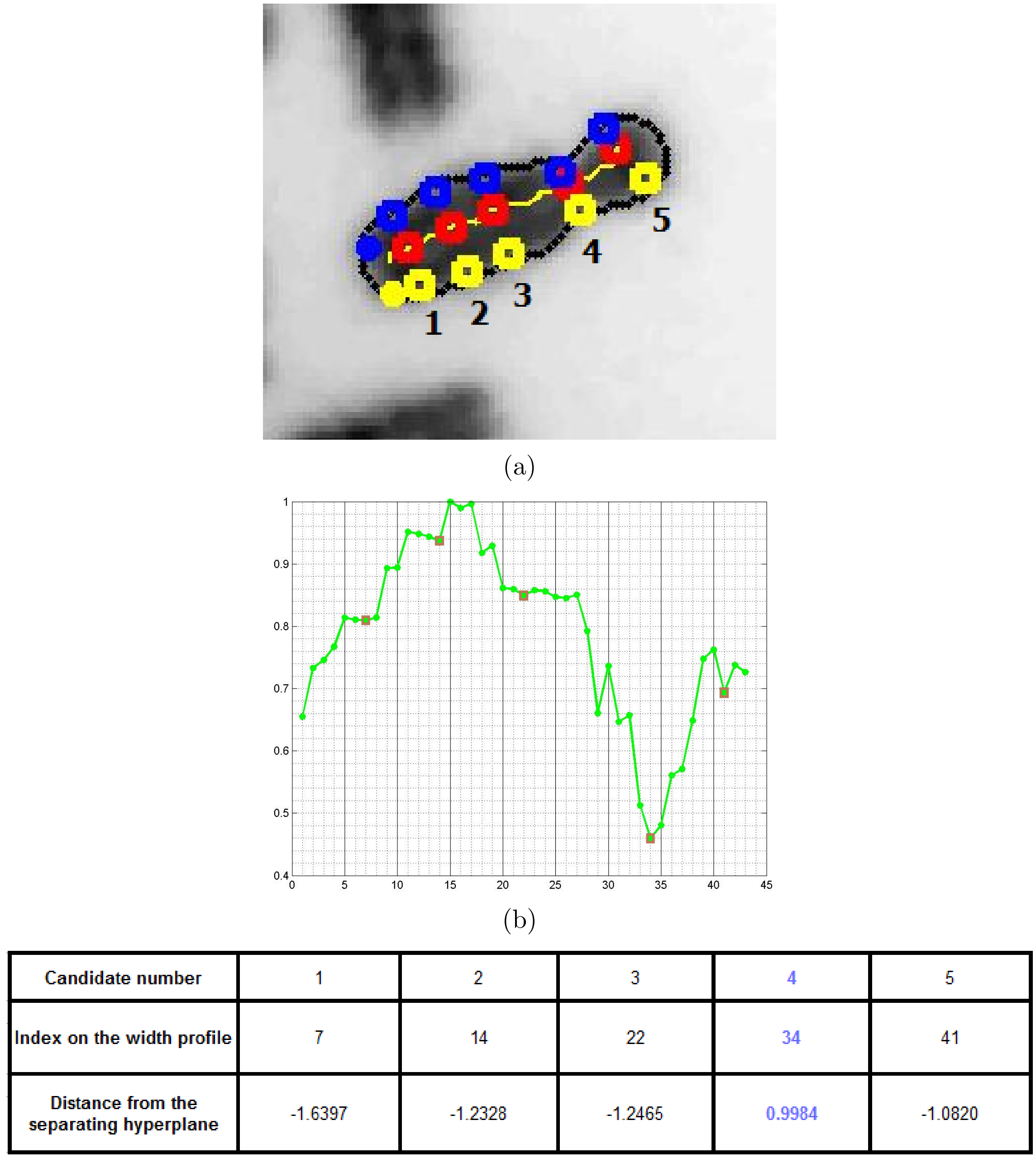
Table 4.11: Breakdown of chromosome cell images and chromosomes used for the larger data set based on the staining method and the sister chromatid separation (SC Sep.)

Abbr.	**Label**	**Images**	**Chromosomes**
D-NSC	DAPI-No SC Sep.	4	114
D-WSC	DAPI-With SC Sep.	18	587
G-WSC	Giemsa-With SC Sep.	18	699
	Total	40	1,400

The centromere locations manually recorded by the author were used as the 'ground truth' in the analysis. The set of candidates generated by the algorithm was displayed superimposed on the chromosome and the candidate(s) that closely represent (within the centromere region) the centromere location was selected while rejecting others. In cases where all the candidates provided by the algorithm are incorrect, all the candidates were marked as rejected (negative examples for the classifier). Since the centromere is a region, a pixel error in detection may not convey the accuracy of the algorithm effectively. Therefore a binary detection accuracy measure was used for this test. In the current stage of the research, the intra-observer variability of ground truth was not analyzed due to the limitations of resources. The 1400 chromosome data set yielded 7058 centromere candidates. A randomly selected 50% portion of the data set along with the corresponding ground truth were used for training a support vector machine for centromere localization. A Gaussian radial basis function kernel was used with sequential minimum optimization (which gives a l-1 norm soft margin classifier) for training the support vector machine classifier in this experiment. The trained SVM classifier was tested for effectiveness using the remaining 50% of the data set (2 fold cross validation) and obtained an accuracy, sensitivity and specificity values of 92%, 96% and 72% respectively. Two fold cross validation was selected as the validation method since it is a less computationally expensive method compared to methods such as the 'leave on out' approach. Furthermore, the importance was placed on the ranking of the candidates as opposed to the label given for each candidate. Therefore two fold cross validation method yielded a reasonable estimation of the performance with minimal computation.

However, the key objective was to accurately detect the centromere location for each given chromosome in the data set as opposed to classifying all candidate points individually. The candidates in each chromosome were analyzed separately and the best candidate from the set was selected based on the distance metric value (ρ). After testing on 1400 chromosomes, the algorithm accurately located the centromere location in 1220 chromosomes with a detection accuracy of 87%. Its also important to note that the 124 chromosomes out of the missed 180 chromosomes were cases where none of the candidates included the centromere of the chromosome. Some of these were caused by segmentation of acrocentric chromosomes where the lighter intensity of the satellites were segmented out while others were caused by extreme sister chromatid separation. The detection accuracy for each image group is given in table 4.12 where a slight reduction in accuracy was observed for the groups with the presence of sister chromatid separation. The lowest detection accuracy was observed with Giemsa stained images which generally shows higher degree of chromosome boundary noise. It is important to notice that these observations are consistent with the conclusions derived through the Games-Howell post-hoc test (see table 4.9) discussed in section 4.1.1.

Figure 4.4 shows an example where 5 candidates were created using the local minima locations in width profile given by the proposed Laplacian based thickness measuring algorithm. In this instance, the 4th candidate was selected which was the largest positive distance from the data set, yielding a truncated CBCC value of 1.00. Figure 4.5 provides a sample representation of cases where the centromere was accurately localized. From a machine learning point of view, figure 4.5 (a), (b)



(a)

(b)

Candidate number	1	2	3	4	5
Index on the width profile	7	14	22	34	41
Distance from the separating hyperplane	-1.6397	-1.2328	-1.2465	0.9984	-1.0820

(c)

Figure 4.4: Demonstrates an example where 5 candidates were created for the chromosome in figure 4.4 (a) using the width profile in figure 4.4 (b). The figure 4.4 (c) shows the signed distance values for each candidate calculated from the separating hyperplane while the selected candidate is depicted in blue.

and (c) are fairly straight forward centromere localizations. The very high truncated CBCC values at 1.000 for all three cases provide further validity into the CBCC measure which indicate that the selected candidate was more preferable than the other candidates in the chromosome. Figure 4.5 (e) represents a chromosome where sister chromatid separation has had a significant effect on the chromosome segmentation. However as a result of correcting for sister chromatid separation, the algorithm has localized the centromere accurately with a CBCC value of 1.000. The chromosome segmentation in figure 4.5 (d) demonstrates evidence of extensive sister chromatid separation and therefore the CBCC value was at 0.995 which still was a high value for the data set. The figure 4.5 (f) represents a chromosome which was highly bent and also with very significant sister chromatid separation present within. Yet, the algorithm was capable of localizing an accurate centromere location with a low CBCC value of 0.661, which indicated a less than ideal separation between the centromere candidates.

Table 4.12: The detection accuracy values for chromosomes used for the larger data set based on the staining method and the sister chromatid separation (SC Sep.)

Abbr- -viation	**Number of chromosomes**	**Number of accurate detections**	**Accuracy**
D-NSC	114	104	91.2%
D-WSC	587	517	88.1%
G-WSC	699	599	85.6%

Figure 4.6 provides some cases where the algorithm failed to localize the accurate centromere location. Most of these (68%) were observed to be cases where none of the candidates were deemed to contain the actual centromere location, mainly due segmentation problems and very high levels of sister chromatid separation. Figure 4.6(b) depicts an example where the segmentation algorithm failed to capture the constriction in an acrocentric chromosome. The CBCC value in this example was as low as 0.066 which indicated that the algorithm picked a weak candidate for the centromere. Figure 4.6(a) demonstrates a case where extreme sister chromatid separation has caused the segmentation algorithm to treat each individual chromatid arm separately. This chromosome had a low CBCC value of 0.368 which represented the acentric nature (morphological) of the separated arm. Another adverse impact of high sister chromatid separation is given by figure 4.6(c) where the long arm sister

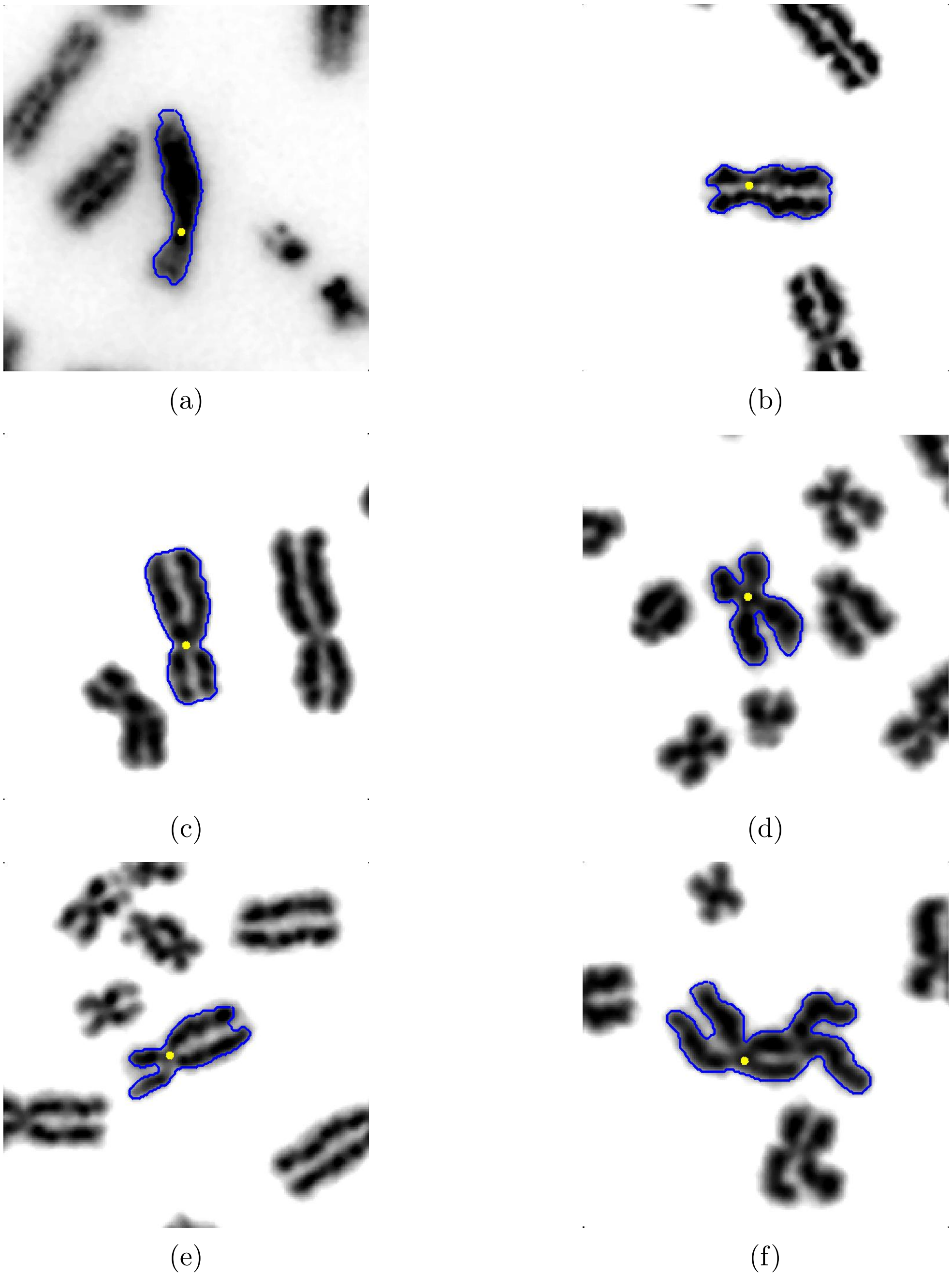


Figure 4.5: Demonstrates some sample results of the algorithm where the accurately detected centromere location (selected candidate) is depicted by a yellow dot while the segmented outline is drawn in blue. Figure 4.5 (a) is a result of DAPI stained chromosomes while figure 4.5 (b)-(f) are results of Giemsa stained chromosomes. These results reported CBCC measures of (a) 1.000, (b) 1.000, (c) 1.000, (d) 0.995, (e) 1.000, (f) 0.661 respectively.

chromatids of an acrocentric chromosome had been identified as a bent chromosome with no sister chromatid separation. The CBCC measure failed to distinguish this chromosome from a normal bent chromosome and had yielded a relatively high (compared to other misidentified localizations) value of 0.655.

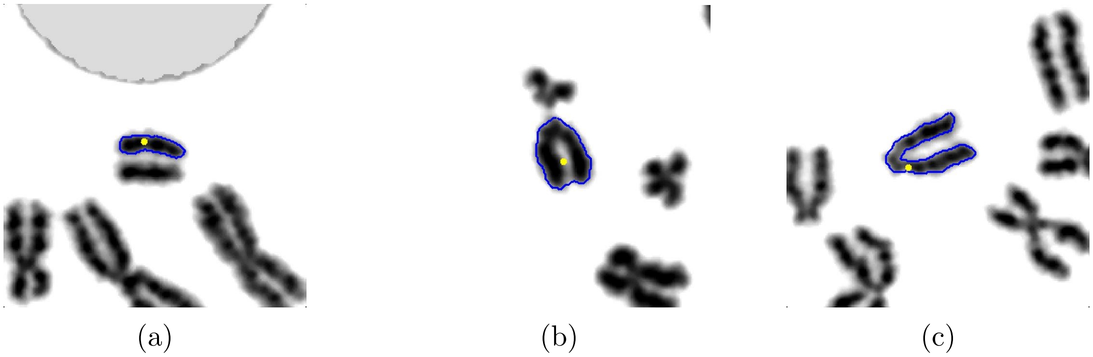


Figure 4.6: Demonstrates some sample results where algorithm failed to yield an accurate centromere location. The detected centromere location (selected candidate) is depicted by a yellow dot while the segmented outline is drawn in blue. These results reported CBCC measures of (a) 0.368, (b) 0.066, (c) 0.655 respectively.

A preliminary study was conducted to gauge the possibility of extending the proposed centromere detection algorithm into dicentric (chromosomes with two centromere locations) detection in radiation biodosimetry. Given that the constriction at the second centromere carries similar characteristics to the first centromere location, in theory it should be ranked high along with the best candidate (primary centromere). Therefore, the top four ranked candidates of the dicentric chromosomes in the data set was analyzed manually. The purpose was to find out whether both centromere locations would be encompassed within the top four candidate positions. In all 31 dicentric chromosomes in the data set, the first candidate (the selected centromere) was accurate. Out of the 31 cases, there were only two instances where the second centromere was not within the top four candidates. This was caused mainly by high sister chromatid separation. The example given in figure 4.5 (f) was observed to be one of these cases. The breakdown of the candidate numbers which captured the second centromere location is given in table 4.13, where a majority of cases reported the second centromere location as the second highest ranked candidate location. It is important to notice that in some of the cases, more than one candidate was created at the primary centromere location (in long chromosomes). This was observed to cause some of the cases where the second centromere was ranked as the third candidate.

Table 4.13: The results of the preliminary analysis in studying the feasibility of extending the proposed method for dicentric detection is presented by indicating the number of times different ranked candidates were able to encompass the second centromere.

Rank of the second centromere	Number of cases
02	20
03	6
04	3
05	1
06	1

Chapter 5 Conclusions & Future work

The dissertation presented a novel algorithm for effectively analyzing human metaphase chromosomes in lymphocyte cell images. The algorithm was tested for the accuracy in width profile calculation as well as for centromere detection accuracy as discussed in chapter 4. This chapter provides a summary of the algorithm along with some conclusive remarks and feasible future work.

5.1 Summary of the proposed method

The algorithm first segmented the chromosome using a multi stage local segmentation algorithm. The segmented object along with the gray scale image was used to reliably calculate the centerline of the chromosome despite the morphology of the chromosome. The centerline calculation process was able to guarantee the presence of no spurious branches in the final result. This provided the algorithm a basis for calculating the length of a given chromosome. Next, the proposed algorithm autonomously selected salient points to partition the contour of the chromosome. This was needed in order to isolate the telomere regions of the chromosome which capture the evidence of sister chromatid separation. This was followed by feature extraction process based on functional approximation which was then used to detect whether the chromosome contour contained evidence of sister chromatid separation or not. The objective of testing for sister chromatid separation was to correct the location where the chromosome contour was split into two approximately symmetrical segments. This correction was essential since the centerline tend to deviate into one of the chromatids in the presence of sister chromatid separation. The method next used a Laplacian based thickness measurement algorithm to calculate the width profile of the chromosome along the longitudinal axis of symmetry. Intensity was integrated in to this process in order to guide the thickness measuring trace lines to closely follow homogenous regions of the chromosome (chromosomal banding). The width profile was then used to calculate a set of candidates for centromere location of the chromosome. Next, these candidates were ranked based on the geometric margin with respect to the separating hyperplane of a trained classifier, where the best candidate was selected as the centromere of the chromosome.

A confidence measure termed as the 'Candidate Based Centromere Confidence' (CBCC) was proposed which conveyed the confidence of each centromere detection using a scalar value with an upper bound of unity. CBCC provides the expert with additional information regarding the detection which they can use to make informed diagnosis.

The proposed algorithm was tested at two different stages for accuracy and performance. First, the capability for calculating an accurate width profile was tested using the centromere localization accuracy in comparison with a centerline based approach [1]. A data set of 226 chromosomes were used for the test where the quantified detection error values were subjected to an in depth statistical analysis. The proposed method was proved to have a statistically significant improvement when compared with the centerline based method while performing more consistently across different staining methods and morphologies. The second stage of testing was conducted on a larger data set containing 1400 chromosomes collected from 40 cell images across two staining methods. Here the accuracy of the proposed method was tested using the candidate based approach, where a detection accuracy of 87% was obtained.

The experiments that were carried out to detect the performance of the proposed method, revealed some limitations. Extreme sister chromatid separation was one of the main contributors to these limitations. Each chromatid arm of some chromosomes were segmented as two separate chromosomes due this phenomenon. In some acrocentric chromosomes, extensive sister chromatid separation gave the appearance of a bent chromosome with no sister chromatid separation. Furthermore, in some acrocentric chromosomes, the segmentation algorithm failed to capture the width constriction altogether. This was mainly caused by the relatively light intensity of the satellite stalks (nucleolar organizer regions) of these chromosomes getting segmented out (see figure 4.6 for some of these example situations).

A number of empirically tuned parameters are reported in this dissertation. It was important to establish the robustness of the proposed method with respect to these parameters. Therefore, the parameter values were first tuned using the preliminary data set of 226 chromosomes. Then the same set of values were used for processing the larger data set of 1400 metaphase chromosome. The detection accuracy of 87% suggests that the empirical values used are considerably robust in the presence of a different data set.

5.2 Conclusive remarks

The centromere detection algorithm used for radiation biodosimetry requires being able to handle chromosomes with multiple staining methods and preparation techniques since the data may be generated at different laboratories with different protocols. These variations in preparation techniques result in large morphological variations along with the presence of premature sister chromatid separation caused by exposure to prolonged colcemid time and concentration. The proposed method was observed to perform satisfactorily despite the high morphological variations on cells images from DAPI as well as Giemsa stained images (see section 4.2). Furthermore, the proposed method performed better at calculating width profile of chromosomes with minimal influence from boundary noise than the centerline based approach (discussed in section 4.1.1).

Dicentric chromosomes appear in low frequencies in human metaphase cell images even at considerable radiation levels and become even less frequent in lower radiation dosages. Therefore, it is paramount to include all types of chromosomes in the analysis for dicentric detection. This is a major drawback in methods currently known [61], [62]. The candidate based approach in the proposed algorithm gives the ability to include both acrocentric and submetacentric chromosomes into the analysis. Coupled with the CBCC (Candidate Based Centromere Confidence) metric, the proposed algorithm is geared to provide useful information to the expert involved in the diagnosis process. Its important to notice that though these are essential requirements for radiation dosimetry, they are also desirable properties to have in any chromosome analysis and centromere detection algorithm.

More research is warranted in extending the centromere detection algorithm to accurately identify dicentric chromosomes. The initial experiment performed to analyze whether the top four ranked candidates can capture the second centromere location (discussed in section 4.2) is an important step towards this direction. It was demonstrated that the algorithm captured 29 out of the 31 second centromere locations within the top four candidate pool. This implies that the proposed method provides a good framework for detecting dicentric chromosomes in radiation biodosimetry applications.

5.3 Future work

The proposed method presents a framework for incorporating additional features into the Laplacian based thickness measurement process. The framework was used to integrate intensity information into the thickness measurement process in the research. However, the possibility of incorporating other relevant features into thickness measurement processes in different applications warrants further experiments.

The proposed algorithm currently assumes at the least one centromere location chromosome exists. This is a reasonable assumption in most cases. However, in radiation dosimetry analysis, there is a possibility of encountering acentric chromosomes which have no centromere location within them. Therefore a mechanism needs to be developed in order to reject these cases in the future. Radiation dosimetry analysis can also benefit from an extension of the current algorithm to detect dicentrics which are chromosomes with two centromere locations. An accurate calculation of these dicentric occurrences provides an insight to the extent of radiation exposure a patient have had over the years.

Furthermore there exists the need for an effective method for separating touching and overlapping chromosomes in order to fully automate the proposed algorithm. There exist some methods in literature which attempted to separate these overlapping/touching blobs by detecting the pale path between the objects or by using a model based approach. The implementation of an accurate algorithm for this problem will ensure the smooth and effective operation of the proposed algorithm by removing the need to select individual chromosomes as well as the need for using a thresholding factor for separating barely touching chromosomes.

In radiation dosimetry, the algorithm needs to process a large number of chromosomes within a short period of time. This is specially the case in a mass casualty nuclear event. Therefore the processing time of the proposed method needs to be reduced in order to facilitate the said requirement. Fast optimization methods such as dynamic programming could be used to speed up stages of the proposed algorithm.

During this study, the content based ranking algorithm [46] was observed to perform poorly on cell images with premature sister chromatid separation. Since poorly ranked images are discarded from the analysis, this can have significant impact on the accuracy of the algorithm and therefore warrants further improvements.

A high centromere detection accuracy is critical for radiation biodosimetry applications since the frequency of dicentric occurrence in cell images are relatively low even at high radiation doses. Therefore, the detection accuracy of the proposed method needs to be improved by implementing a better segmentation algorithm which can reliably capture the centromere constriction of acrocentric chromosomes. A mechanism for excluding cell images with extreme levels of sister chromatid separation from the analysis will further contribute to improve the detection accuracy.

The proposed algorithm can be tested on a larger data set containing other staining methods such as G-banding in order to gauge the effectiveness of the algorithm. Since the author collected the ground truth in this experiment, it is important to involve an expert to annotate a separate data set to verify the performance measures reported in this dissertation. The ADCI software which is currently on field testing could provide the perfect platform for collecting the required ground truth data. It is also important to test the inter and intra observer variabilities during the analysis of such a data set.

Appendices

Appendix A Background methods

A.1 GVF snakes

Gradient vector flow (GVF) snakes is a widely used active contour model in segmentation. It is a well known method in literature that has being used to obtain better convergence at boundary concavities. Section A.1.1 provides a brief overview of the traditional active contours and section A.1.2 discusses the GVF snake external energy model in detail. A brief comparison between the GVF snakes and Distance Transform (DT) based snakes is stated in section A.1.3.

A.1.1 Parametric snakes or active contours

In 1988, Kass et al. [29] first introduced parametric active contours and they have been applied to many image processing problems ever since. This approach can be modeled either as an open or closed curve within the 2D domain of the image where the contour iteratively deforms in order to conform to image features such as edges [63]. A parametric curve (PC) in general, can be stated as in equation A.1.

$$v(s) = (x(s), y(s)) \qquad 0 \leq s \leq 1 \tag{A.1}$$

In order to achieve this behavior, the curve is either shrunk or expanded based on the value of the internal energy term defined for the curve. Convergence occurs when this internal energy term is neutralized by an external energy term (also known as "data term") acting upon the curve at that specific position. Therefore, the energy formulation of the snake model can be viewed as an energy (physics based) minimization problem depicted by equation A.2.

$$E_{snake} = \int (E_{internal}(v(s)) + E_{external}(v(s)))ds \tag{A.2}$$

In order to represent the contour as a set of 2D control points, the energy terms in equation A.2 can be modified and converted into the discrete domain as given by equation A.3 and equation A.4.

$$v_i = (x_i, y_i) \qquad 0 \leq i < n \tag{A.3}$$

$$E_{snake} = \sum_{i=0}^{n-1} (E_{internal}(v_i) + E_{external}(v_i)) \tag{A.4}$$

The internal energy ($E_{internal}$) stated in equation A.4 is a combination of two characteristics which govern the motion of the snake when it is under no influence from the data term. Equation A.5 (by discrete approximation) below depicts these two characteristics respectively as elasticity and stiffness and they ensure that the evolution of the contour under the internal forces does not deform the original shape. The constants α and β are the corresponding scaling factors which need to be set depending on the application in order to decide the contribution of each energy term towards the motion. Setting these parameters is critical in getting a proper segmentation result. For an example. a higher value of α and β can overpower the external energy component and continue the evolution past the object boundary.

$$E_{internal} = \sum_{i=0}^{n-1} \alpha|v_{i+1} - v_i|^2 + \beta|v_{i+1} - 2v_i + v_{i-1}|^2 \tag{A.5}$$

The external energy in equation A.4 was originally defined to repel the motion enforced by the internal energy component near object boundaries and to incorporate the edge information of the image. This interpretation is shown in equation A.6 where the term $\nabla I(v_i)$ defines the edge strength at control point v_i of the image I. Even though parametric snakes are in general solved as an energy minimization problem, for the external energy (data term energy), it is desirable to maximize this value. Therefore a negative sign is used in equation A.6 to correspond with the general energy minimization framework. Furthermore, γ is the scaling factor used to balance the external and internal energy to prevent the snake from missing edge points and ultimately shrinking to a single point.

$$E_{external} = -\sum_{i=0}^{n-1} \gamma |\nabla I(v_i)|^2 \tag{A.6}$$

Reconsidering the parametric representation (equation A.2), it can be further shown that a snake which minimizes energy should satisfy the Euler-Lagrangian equation shown below.

$$\alpha v''(s) \;-\; \beta v''''(s) \;-\; \nabla E_{external} \;=\; 0 \tag{A.7}$$

By treating $v(s)$ as a function of time, this can further be expressed as a dynamic equation with (t) [30],[64]:

$$v_t\,(s,\,t) \;=\; \alpha v''(s,\,t) \;-\; \beta v''''(s,\,t) \;-\; \nabla E_{external} \tag{A.8}$$

The above equation will be used to represent the GVF snake energy terms which will be discussed in the subsequent section A.1.2.

Parametric active contours defined above had been applied to many segmentation problems in a variety of fields. The main advantages of this model are listed below.

- As opposed to methods based on edge maps, they yield a connected contour as the end result .
- Active contour model evolves under the influence of cumulative forces on all control points in contrast to point processing methods such as thresholding.
- When segmentation problem is modeled as an energy minimization problem, it can effectively be solved mathematically.

Yet, the basic snake model described by Kass [29] has the following disadvantages,

- Snakes are not guaranteed to find the global solution for the problem and depending on the initialization they often converge to a local minima .

- Basic parametric active contour model cannot handle topological changes in the object of interest. Therefore the number of objects in a given image need to be predetermined to calculate the object boundaries.
- Snake contours have the possibility to twist (fold on each other) which is highly unlikely to be present in real objects.
- The model has very limited capture range for the data term and thus has problems in negotiating concave boundaries.

Among these limitations in the traditional snake model, the sensitivity to the initialization and the low capture range have been identified as most prominent. The most common method used to increase the capture range is by merely spreading the edge strength using Gaussian smoothing. The Gaussian filtering increases the range in which the snake movement can be influenced by the edge strength. This can be represented as in equation A.9. Yet, depending on the variance (σ) value of the Gaussian distribution used [65], this application also blurs the image boundaries and the exact positioning of the end segmentation result . The capture range will be larger, if the variance (σ) is set to a higher value, but this will also produce highly blurred image boundaries which can adversely affect the accuracy of the final contour positioning.

$$E_{external} = -\sum_{i=0}^{n-1} \gamma |\nabla(G_\sigma * I(v_i))|^2 \tag{A.9}$$

Another common approach to increase the capture range is to apply the standard distance transform (DT) to the image of interest. In order to apply this transform, a distance map is created using intensity edges as the feature points where the distance value is set to be proportional to the shortest distance from any of these feature points. Equation A.10 illustrates the use of distance transform as an external energy in parametric snakes. Unlike the external energy functions given by equation A.6 & A.9, there is no negative sign in equation A.10 because the value of $D\left(v_i\right)$ reduces as the point becomes closer to image boundaries (feature points).

$$E_{external} \ = \sum_{i=0}^{n-1} D\left(v_i\right) \tag{A.10}$$

The distance transform $(D\,(p))$ for any pixel p in the image can be formally defined as following,

$$D\,(p) \;=\; min_q\;\{\alpha.\,\|p\;-\;q\|\;+\;F(q)\}$$

where $F(.)$ is a modified 2D matrix generated based on the feature points or the edge map. For 'standard distance transform', given the edge map $edge(I)$, the function $F(p)$ for any pixel p can be defined as follows ,

$$F(p) \;=\; \begin{cases} 0\;, & \text{if } p \in edge(I) \\ \infty, & \text{if } p \notin edge(I) \end{cases}$$

In order to further enhance the capture range of a snake, Cohen [66],[67] applied a non-linear transformation to the above mentioned distance maps . These 'distance potential forces', only altered the magnitude of the potential forces which are acting on the edge map while retaining the original vector filed orientation.

A.1.2 Gradient vector flow as an external energy

As mentioned above, there are two main limitations in using conventional snake model in real world segmentation problems. High sensitivity to the initialization of the snake control points with respect to the data terms present in the image can be considered as one limitation. For example, if the adopted model is of a shrinking snake and the initial contour is selected completely within the object boundary, the snake would evolve into a single point missing the actual boundary. At boundary concavities, the direction of the image gradient (on each side of the concavity) would point in opposite direction and prevents the snake from converging toward concave regions (see figure A.2). Therefore, given the edge map $(edge(I))$ such that,

$$edge\left(I_{x,y}\right) \;=\; E_{external}\,(x,y)$$

we can define a static vector field $\mathrm{v}\,(x,y) \;=\; [u\,(x,y)\,,\,v\,(x,y)]$, which minimizes the energy functional [30] given below,

$$\varepsilon \;=\; \iint\limits_{x,y} \mu\left(u_x^2 \,+\, u_y^2 \,+\, v_x^2 \,+\, v_y^2\right) \,+\, |\,\nabla edge\,|^2\,|\,\mathrm{v} \,-\, \nabla edge\,|^2 \; dxdy \qquad \text{(A.11)}$$

where $\nabla edge$ is the gradient of the edge map and u_x is the partial derivative of component $u(x, y)$ with respect to x ($u_x \;=\; \frac{\partial u(x,y)}{\partial x}$). By closely observing the equation A.11, following two behaviors of the energy functional ε can be revealed,

- At homogenous regions where $\nabla gradient$ is small , the functional ε is influenced by the partial derivatives of the vector field, thus ensuring a smooth variation along the homogenous regions. Traditional energy models would have no response to such behavior. Therefore, this first term of equation A.11 is called the "smoothness term", where μ is the factor used to balance the contributions from the two terms towards ε.

- In regions close to object boundaries, the $\nabla gradient$ value becomes more dominant increasing the contribution of the second term of equation A.11. The minimum value (0) for the energy functional in the same equation is achieved by setting v $=$ $\nabla gradient$ around the vicinity or object boundaries preserving the conditions for a fast convergence.

Then by replacing the external energy component ($E_{external}$) in equation A.8 with v (obtained by minimizing equation A.11), the following representation can be achieved,

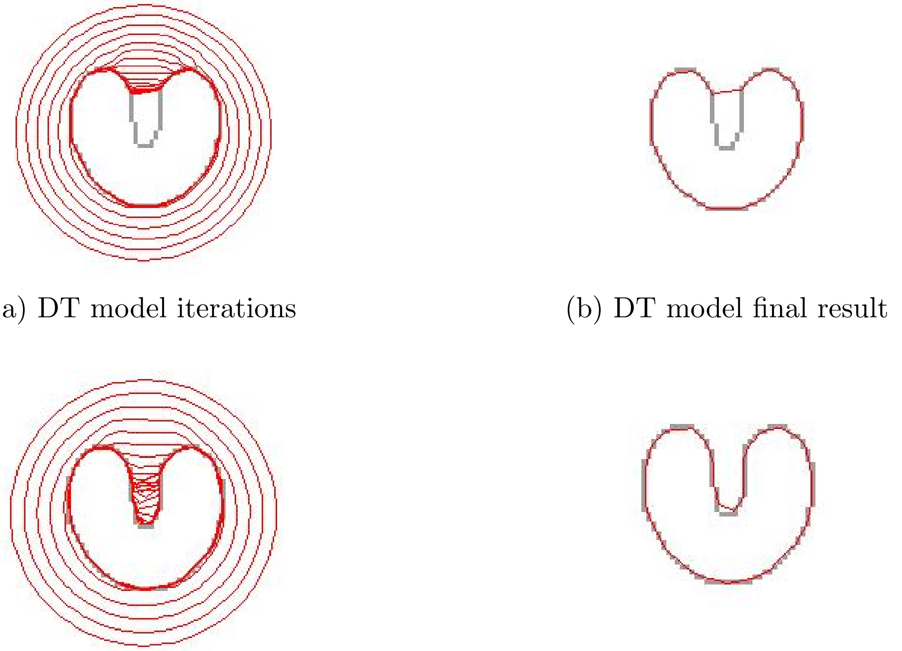
$$v_t\,(s,\,t) \;=\; \alpha v''(s,\,t) \,-\, \beta v''''(s,\,t) \,-\, \mathrm{v} \qquad \text{(A.12)}$$

GVF snake is the parametric curve obtained by solving the equation A.12. Detailed information on solving this equation can be found in [30],[64],[65].

A.1.3 GVF snakes vs DT snakes

In this section, image results of applying two different external energy models to active contours are examined, based on the image and code examples provided by Prince & Xu [5]. Figure A.1 depicts the difference in convergence of the DT based snake with the GVF snake. Figure A.1(a) and A.1(c) show that the GVF converges faster and more deeper into the concave region of the image where as the final result of the DT snake (figure A.1(b)) is not satisfactory. The reason for GVF snake to converge into the region where DT snake fails, can be explained by using the respective vector fields given in figure A.2. Following observations can be made with respect to the two vector fields,

- The GVF model vector field is more dense relative to the DT snake field and is specifically stronger near object edges. It decreases (in magnitude) slower than the DT model, when going away from these boundaries, thus explaining the faster convergence.
- The GVF field points towards the concave boundary (in the mid section of the U shape concavity) whereas the DT model vectors simply exert forces with opposite directions (in the same region). This cancels out the influence of the external energy term in the DT model and therefore, the snake stops traversing towards the concave region.



(a) DT model iterations

(b) DT model final result

(c) GVF model iterations

(d) GVF model final result

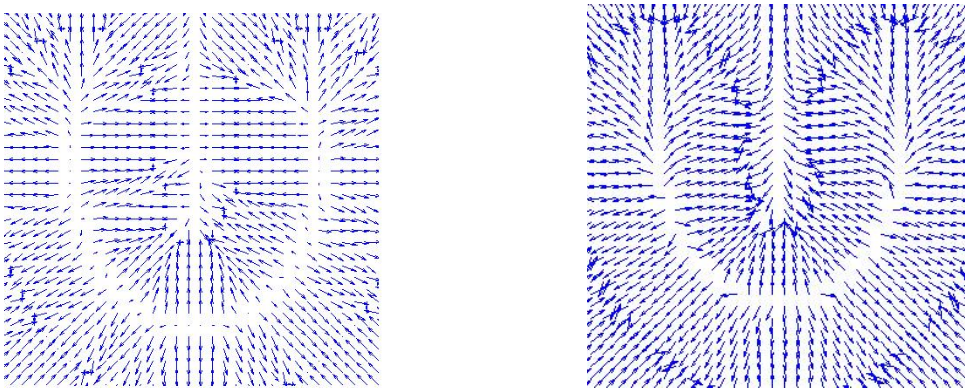
Figure A.1: Comparison between Distance Potential (DT Based) model (top) and the GVF model (bottom). Each model depicts the initialization of the contour and convergence with each iteration (on left) followed by the final contour result after 100 iterations(on right) [5]. (Reproduced with permission from Prof. Jerry L Prince)

A.2 Discrete curve evolution (DCE)

In many shape matching applications, it is required to compare multiple objects variations with an underlying shape structure. DCE is a technique that can be adopted for evolving polygons to preserve visual information which can yield a hierarchical set of polygons according to their significance in representation of the original object. It is an effective and robust tool for generalizing polygonal contours based on digital linearization [6]. This technique can be directly applied to digital images as the boundary of any digital image object can be approximated with a polygon containing high number of vertices.

This contour evolution method is observed to have potential applicability in the following fields of studies:

- Shape simplification : DCE can be directly used for shape simplification to



(a) DT vector field

(b) GVF vector field

Figure A.2: Comparison between Distance Potential (DT Based) vector field (left) and the GVF vector field (right) [5]. (Reproduced with permission from Prof. Jerry L Prince)

compare different shapes of objects derived from the same model. Latecki & Lakämper [50] have done detailed study of this aspect in one of their publications. Some shape simplified examples can be found on web resources [68]

- Object extraction: DCE can also be used to extract objects from a database when a query is given in the form of a visual sketch. Here, a shape descriptor can be extracted using a simplified polygonal contour obtained through DCE (with a single contour). A study on this application on the MPEG-7 standard data set was performed by Latecki et al [69].

- Skeleton pruning: Bai et al [49] suggest that the polygons which are obtained through the DCE process can be effectively used for skeleton pruning as it can remove boundary noise from the digital image object. Also, the DCE end result would be a high level representation (based on the relevance function) of the initial object. During our research, we have used this application of DCE to obtain the centerline of a chromosome.

A.2.1 Definitions

First, we will briefly define the notations that will be used in this sections to explain the DCE process. Let, $C \in \Re^2$ be the contour of interest which may also contain self-intersections. Then we can define P as a closed polygon which will lead to a sequence of polygons $\left(P^0,\, P^1,\, ...,\, P^{m-1},\, P^m\right)$ through DCE. Also, we can define the following general terms (related to the polygon structuring),

- $v\left(P^i\right)$ as a vertex contained in the polygon P^i
- $arc\left(s_i,\, s_{i+1}\right)$ defines the arc that spans between two line segments as $s_i \cup s_{i+1}$
- the line segment s_i consists of a line connecting two adjacent vertices and is defined as $v_i \,\cup\, v_{i+1}$

Then we can define a relevance value $K(v,\, P^i))$ for any vertex on the closed polygon P. section A.2.2 explains the equation used for calculating this relevance measurement and the rationale behind using it. The algorithm of discrete curve evolution by digital linearization is illustrated below [6],[52],[70],

The DCE algorithm:

1. Find the value of,

   $K_{min}\left(P^i\right) \;=\; min\left\{K(u, P^i) \mid u \in v(P^i)\right\}$, where (K_{min}) is the minimum value for the relevance measurement at a given iteration.

2. Find the set (V_m) which give all the vertices with the minimum relevance value found above(K_{min}) and this can be noted as,

   $$V_m\left(P^i\right) \;=\; \left\{u \,\in\, V(P^i) \;\mid\; K(u,\, P^i) \;=\; K_m(P^i),\, \forall\, i = 0, 1, ..., (m-1)\right\}$$

3. Then, DCE is the process of constructing a new polygon P' from the previous P polygon by deleting all the vertices with the minimum relevance value (K_{min}). This can be expressed as,

   $$V\left(P^{i+1}\right) \;=\; V\left(P^i\right) \,\setminus\, V_m\left(P^i\right)$$

   where $|V\left(P^m\right)| \;\leq\; 3$, in which $|\,.\,|$ is the cardinality operator

This new polygon is created by replacing two line segments ($s_i \,\&\, s_{i+1}$) with a new line segment s', which effectively connects the end points of $arc\,(s_i,\, s_{i+1})$ provided that the arc has a relevance value of K_{min}.

The iterations can be carried out for any desired termination criterion even until the end polygon becomes convex.

Depending on the application when the end criterion is set up accordingly, the end result convexity can be assured by stopping the process at a higher stage of evolution. Since the convex shapes determines the visual parts of an object [6], the convexity of the result polygonal partitions are of utmost importance . If the stopping criterion is inappropriate, the algorithm will converge to a degenerate solution of a polygon $P = \{\varnothing\}$. The shape simplification process and the immunity to noise of the above defined function can be clearly seen in figure A.3[1]. Some feature points are marked in figure A.3 to show the stability with noise deformations and the similarity of the two evolution results.

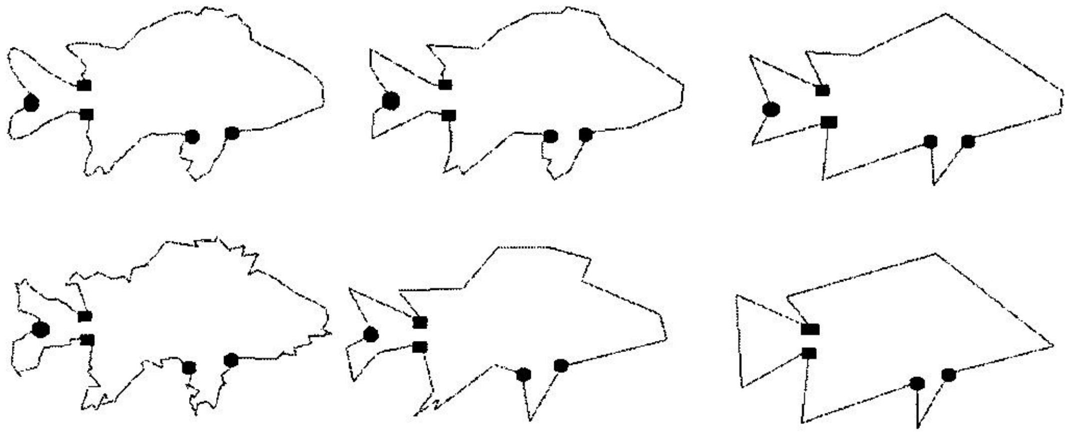


Figure A.3: above: Steps of the DCE process of shape simplification, below: the same steps when initiated by adding boundary noise to the same image [6]. (Reproduced with permission from Prof. Longin Jan Latecki)

[1] Special acknowledgement to Prof. Longin Jan Latecki, Dept. of Computer and Information Sciences, Temple University, Philadelphia, USA.

A.2.2 The relevance function

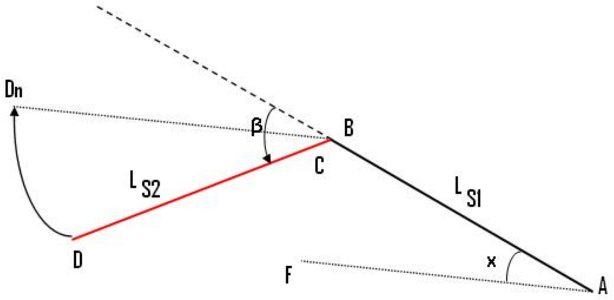
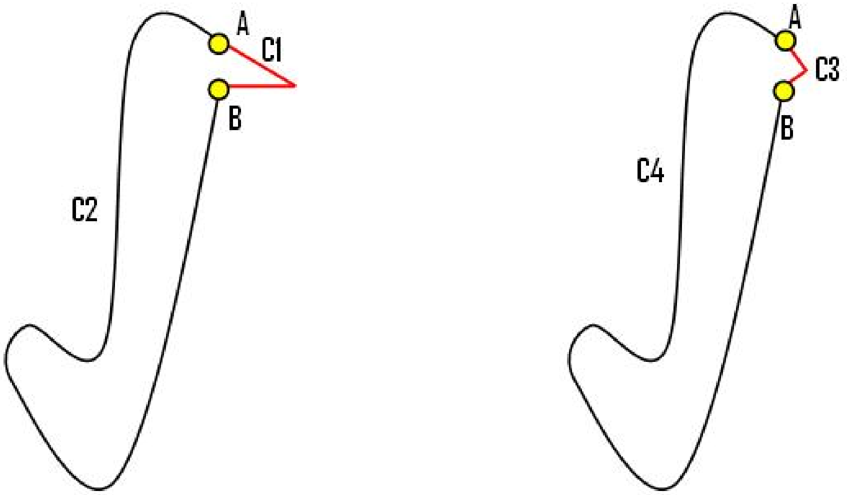


Figure A.4: A representation of two line segments and the used angle measurements in the DCE process, which can be used to explain the rationale behind the used 'relevance measure'.

Discrete curve evolution is used to obtain a hierarchical set of polygons that represent the shape features of the original contour $C \in R^2$. The effectiveness of this evolution process depends on the measurement used to select vertex/vertices to be deleted at each iteration in order to obtain a better and simpler representation of the original object. The main assumption behind deriving this relevance equation (see equation A.13) is as follows:

'Larger values of total turn (arc) angles as well as relative lengths of segments imply higher contribution to the shape of the curve or in other words, these segments have higher relevance value'

The highly intuitive rationale behind the above assumption can be explained using figure A.5. In this figure line segments A-B & C-D are two segments (B = C) on a polygon partition (P^i) in the DCE process, which makes the arc ($arc\,(s_1,\, s_2)$). Then the turn angle as shown in figure A.4, is expressed as $\beta\,(s_1,\, s_2)$ and is calculated by $|angle\,(C\,-\,D)\,-\,angle\,(A\,-\,B)|$. According to figure A.5, the contour segments $C2$ and $C4$ are equal in length and shape whereas those of $C1$ and $C3$ arcs are not. The shape contribution of the arc $C1$ is higher than that of $C3$ with respect to rest of the contour. Also, the turn angle and the length of the segments of $C1$ is greater than $C3$. Therefore, these observations justify the assumption made earlier regarding the relevance measure dependencies. During this process, the lengths of the segments are normalized with respect to the total length of the contour/polygon in order to get the global perspective.



(a)

(b)

Figure A.5: Shape variations of polygon partitions and the effects of turn angle and arc length to the relevance measure.

Next step is to formulate a suitable function to reflect the previously mentioned two parameters. Considering the 'tangent space' representation of the polygon where the x and y axes represent the segment length (normalized) and the direction of each segment respectively, the turn angle (β) is the difference of 'y' axis values between two consecutive entries. Then, an angle $\measuredangle x$ ($0 \leq x \leq \beta$) is calculated, which is the angle that the segment $C - D$ has to be rotated so that point D and D_n coincide where $A - F$ and $B - D_n$ are parallel to each other (refer figure A.4). This angle $\measuredangle x$ can be expressed as $\measuredangle x = [\beta\,(s_1,\, s_2) \times L_{s2}]\ /\ [L_{s1} + L_{s2}]$ where both L_{s1} and L_{s2} are normalized lengths of the segments [6]. Then, the circular arc-length ($L_{s1} \times x$) is defined as the relevance function for the DCE process as follows,

$$K\,(s_1,\, s_2) = \frac{\beta\,(s_1,\, s_2) \times L_{s1} \times L_{s2}}{(L_{s1} + L_{s2})} \tag{A.13}$$

For a given arc of the polygon, the above relevance value can be considered as the linearization cost . This relevance measure, although its calculated locally has a global representation due to normalization of length values. Further explanation regarding the derivation of the equation A.13 and the tangential space representation can be found in [6] along with some image examples at [68].

A.2.3 Advantages & disadvantages of DCE

The discrete curve evolution based on digital linearization has shown the following advantages compared to other existing methods for shape simplification [6], [52], [70] (One of the main comparison methods is the shape simplification work carried out by Siddiq and Shokoufandeh [71]).

- DCE method is rotation, reflection, translation and scaling invariant. It is rotational invariant due to the use of the tangent space for the polygon evolution process.

- DCE is robust in real world discrete digital images, unlike other methods which are based on local extremal points. It is also robust against boundary noise (digitization errors) in digital images by removing the noise in the early stages of the evolution itself. Therefore, the evolved contour is noise free after few initial iterations. The continuity of the DCE method has proven to be stable against noise. [70]

- The algorithm is guaranteed to converge as at least one vertex is deleted in every iteration.

- DCE is guided by a global feature called the relevance value. Though this feature is locally calculated for each vertex pair, it is formulated in a way to represent the contribution of a given arc with respective to the whole contour.

- Throughout the evolution process the relevance value of the polygon to its original shape, reduces gradually. This yields a relevance hierarchy of polygonal representations.

- The DCE method can handle self intersecting objects, objects with holes as well as any object with a complex shape as long as it is possible to obtain a rough approximation for the outer silhouette of the object .

- Since the algorithm operates by merely deleting vertices, it doesn't dislocate the object boundary or blur the boundary as with other methods.

The main drawback of the DCE method is the ambiguity regarding the stopping criteria of the process. A higher level of knowledge of the desired end result (polygon) is a necessity. If not specified, the DCE based method will continue deleting at least one vertex pair in an iteration until the end polygon becomes an empty set.

A.3 Support Vector Machines (SVM)

Support Vector Machines (SVM) a.k.a kernel methods, is a supervised learning method which has its beginning rooted in the Statistical Learning Theory (SLT) [72]. The objective of the technique is simply to generate input-output mapping from a set of labeled training data and hence the term 'supervised learning'. This machine learning technique can be used for classification as well as for regression and is considered the best off the shelf learning algorithm up to date. SVM has attracted popularity in data mining and machine learning for solving real world problems in fields such as bioinformatics, text mining and image analysis. This section is focused on deriving the SVM classifier and discussing some of its properties in detail.

Classification problems have evolved to be much more complicated over the years. Many current machine learning problems or data sets suffer from a phenomena commonly referred to as the 'curse of dimensionality'. This is caused by the increase in the volume as well as the dimensionality of the feature space with the addition of many features to the classification problem. The sparsity of data also increases along with this by scattering the training data points in the high dimensional feature space. Support vector machines, which are based on the optimal margin classifier method, provides an ideal platform for tackling learning problems in high dimensional feature spaces. SVM also provides better generalization compared to other learning algorithms and therefore can perform better with unseen data during testing. Artificial Neural Networks (ANN) can be considered as the other method which can be used to draw some comparison to SVM from a performance point of view. However, SVM demonstrates superior (or comparable) results against ANN solutions due to its inherent properties discussed later in this section [73].

The SVM framework is derived using two main intuitions as described below.

First intuition:

Let $x^{(i)}$ be a given input feature vector and Θ be the training parameter of a classifier. Then y will be the predicted label for the input $x^{(i)}$ based on the following criteria.

Predict $y = +1 \Leftrightarrow \Theta^t x^{(i)} \geq 0$

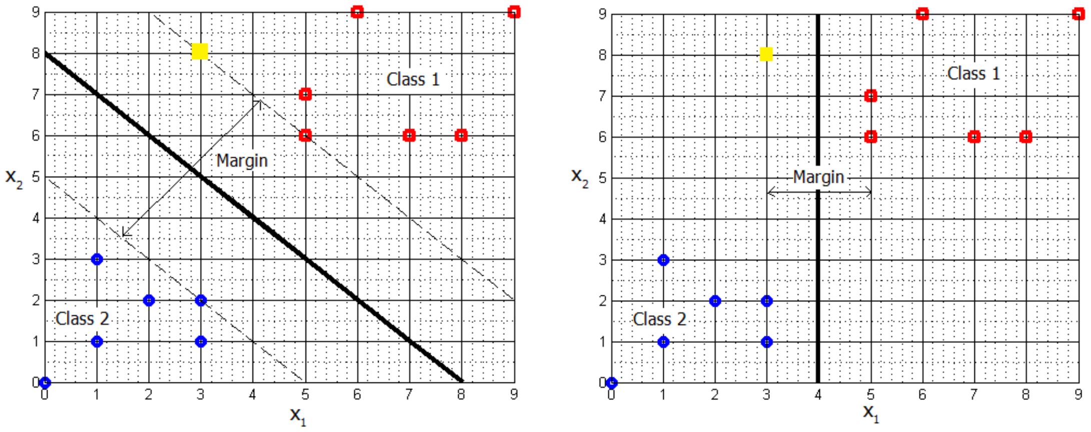
Predict $y = -1 \Leftrightarrow \Theta^t x^{(i)} < 0$

Intuitively, if $\Theta^t x^{(i)} \gg 0$, this implies that the confidence of the decision of assigning the label as $y = +1$ is very high. Similarly, $\Theta^t x^{(i)} \ll 0$ would imply a highly confident label $y = -1$. Therefore we would prefer to see the following situation in a classification problem where,

$$\Theta^t x^{(i)} \gg 0, \; \forall i \;\; s.t \;\; y = +1$$
$$\Theta^t x^{(i)} \ll 0, \; \forall i \;\; s.t \;\; y = -1$$

Second intuition:

Given a classification problem, we would like to obtain the largest margin of separation possible. Figure A.6 depicts a linearly separable training data set where two different separating hyperplanes were derived to yield zero training error. Only the separating hyperplane given by figure A.6(a) contains the optimal margin. This is important in generalizing the classification problem in order to cater unforseen data. The new data point marked by the yellow square is now incorrectly classified in figure A.6(b) due to the small separation margin. Therefore, the second intuition is that we want to maximize this separation margin for both positive and negative samples.



(a) (b)

Figure A.6: Two different separating hyperplanes derived for the same training data points. Figure A.6(a) contains the separating hyperplane with the optimal margin. The new data point given by the yellow square will be classified differently in these two cases.

In the following section, we have briefly followed the steps involved in deriving the SVM algorithm highlighting features which makes the SVM, the best off the shelf machine learning tool in industry up to date.

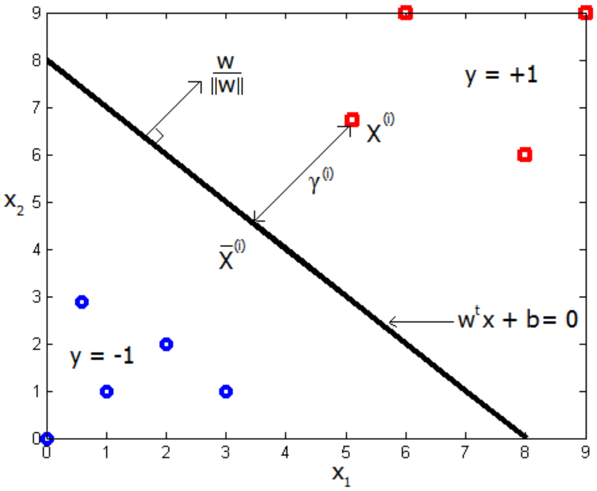
A.3.1 Deriving the SVM framework

Consider a linearly separable classification problem of which the training data points are plotted in figureA.7. Let the labels of the training set be y ($\in \{-1, +1\}$) and the hypothesis (separating hyperplane) h is given by,

$$h_{w,b}(x) \;=\; g(w^t x \,+\, b) \;=\; 0 \tag{A.14}$$

where $g(.)$ is the output function such that,

$$g(z) \;=\; \begin{cases} +1 & if\ z \geqslant 0 \\ -1 & if\ z < 0 \end{cases}$$



(a)

Figure A.7: The framework for deriving the geometric margin of a support vector machine classifier.

The functional margin ($\hat{\gamma}^{(i)}$) of the separating hyperplane given by equation A.14, with respect to a data point $x^{(i)}$ in the feature space is defined as below [74],

$$\hat{\gamma}^{(i)} \;=\; y^{(i)}.(w^t x \,+\, b) \tag{A.15}$$

We would expect positive large values for the functional margin ($\hat{\gamma}^{(i)}$) ideally for all samples in the training data set. A negative value for $\hat{\gamma}^{(i)}$ implies a misclassification. However the value given by equation A.15 can be artificially boosted by scaling up w and b. Therefore, the geometric margin ($\gamma^{(i)}$) which is the normalized distance from the data point $x^{(i)}$ can be derived since it is immune to such artificial boosting. The point $\bar{x}^{(i)}$ which is the projection of point $x^{(i)}$ on to the separating hyperplane should satisfy the equation A.14 (refer figure A.7).

$$\bar{x}^{(i)} \;=\; x^{(i)} \;-\; \gamma^{(i)}\,\frac{w}{\|w\|}$$

Therefore by equating the above coordinates in equation A.14, the geometric margin ($\gamma^{(i)}$) can be derived as follows,

$$\gamma^{(i)} \;=\; y^{(i)}.\left[\left(\frac{w}{\|w\|}\right)^{t} x^{(i)} \;+\; \frac{b}{\|w\|}\right] \tag{A.16}$$

Which implies the general relationship between the two distance measurements given below,

$$\gamma^{(i)} \;=\; \frac{\hat{\gamma}^{(i)}}{\|w\|} \tag{A.17}$$

Furthermore, the geometric margin of a data set (γ) is defined as follows,

$$\gamma \;=\; \underset{i}{min}\,\gamma^{(i)}$$

Therefore, the geometric margin of a data set (γ) which is the worst case for the data set is a good measure to maximize for calculating the optimal margin.

$$\underset{\gamma,w,b}{max}\;\gamma \quad subjected\,to,\;\; y^{(i)}.(w^{t}x^{(i)} \;+\; b) \;\geqslant\; \gamma$$

Using equation A.17 and setting $\|w\| \;=\; 1$ we can obtain the following optimization problem,

$$\underset{\gamma,w,b}{max}\;\frac{\hat{\gamma}}{\|w\|} \quad subjected\,to,\;\; y^{(i)}.(w^{t}x^{(i)} \;+\; b) \;\geqslant\; \hat{\gamma}$$

Then by setting $\hat{\gamma}^{(i)} \; = \; 1$ and changing into a faster minimization problem, the following optimization equation can be derived,

$$\underset{w,b}{min} \;\; \|w\|^2 \quad subjected\,to, \;\; y^{(i)}.(w^t x^{(i)} \, + \, b) \; \geqslant \; 1 \tag{A.18}$$

The equation A.18 present a convex optimization problem which guarantees a global minimum solution and can be solved using techniques such as quadratic programming. However in this form, the computational cost is high for data sets with high dimensionality and large cardinality. Therefore by using Karush Kuhn Tucker theorem [75], the dual problem ($W(\alpha)$) can be derived, where $\alpha_i \geq 0$ is the lagrangian multiplier coefficient for the support vector data point $x^{(i)}$ [76].

$$W(\alpha) \; = \; \sum_i \alpha_i \; - \; \frac{1}{2} \sum_i \sum_j \alpha_i \, \alpha_j \, y^{(i)} \, y^{(j)} \left\langle x^{(i)} . x^{(j)} \right\rangle \tag{A.19}$$

$$w \; = \; \sum_i \alpha_i \, y^{(i)} \, x^{(i)} \tag{A.20}$$

$$\sum_i y^{(i)} . \, \alpha_i \; = \; 0 \tag{A.21}$$

The calculation of w given by equation A.20 reveals that only the support vectors contribute to calculating the separating hyperplane since other α_i are usually equal to zero. Therefore, the support vector machines can handle large data sets efficiently provided that a significant proportion are not support vectors. SVM also utilizes the ”covers theorem” [77] which states that a pattern classification problem that cast into a high dimensional feature space using non linear transformations, is more likely to be linearly separable than in a low dimensional feature space. Through the use of non linear kernels, support vector machine framework transforms the original problem into a high dimensional feature space. The inner product $\left\langle x^{(i)} . \, x^{(j)} \right\rangle$ can be implicitly calculated through the use of the kernel functions [78]. Therefore, SVM provides an efficient framework to classify new data with high generalization compared to other methods.

References

[1] A. Subasinghe A. et al., “An image processing algorithm for accurate extraction of the centerline from human metaphase chromosomes,” in *International Conference on Image Processing (ICIP)*, pp. 3613–3616, September 2010.

[2] M. Y. Karsligil, M. Elif & Karsligil, *Fuzzy Similarity Relations for Chromosome Classification and Identification*, vol. 1689 of *Lecture Notes in Computer Science*, pp. 142 – 148. Springer-Verlag Berlin Heidelberg, January 1999. .CAIP 99.

[3] C. Hilditch, “Linear skeletons from square cupboards,” in *Machine Intelligence*, vol. 4, pp. 403 – 420, Edinburgh Univ. Press, 1969.

[4] L. Lam and S. W. Lee, “Thinning methodologies-a comprehensive survey,” *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, vol. 14, pp. 869 – 885, September 1992.

[5] C. Xu and J. L. Prince, “http://iacl.ece.jhu.edu/projects/gvf.” A web site for the GVF snake demos and examples.

[6] L. J. Latecki and R. Lakämper, “Convexity rule for shape decomposition based on discrete contour evolution,” *Computer Vision and Image Understanding*, vol. 73, pp. 441 – 454, March 1999.

[7] D. Pinkel and J. Landegent, “Fluorescence in situ hybridization with human chromosome-specific libraries: Detection of trisomy 21 and translocations of chromosome 4,” *Proc. Nati. Acad. Sci. USA*, vol. 85, pp. 9138–9142, December 1988.

[8] R. King et al, *A dictionary of genetics.* Oxford university press, fifth ed., 1968.

[9] M. Moradi and S. K. Saterahdan, “New features for automatic classification of human chromosomes : A feasibility study,” *Pattern Recognition Letters*, no. 27, pp. 19–28, 2006.

[10] S. K. Williams and M. R. Cummings, *Concepts of Genetics.* Prentice Hall, fifth ed., 1997.

[11] S. D. Pack and o. A. Stratakis, "Chromosomes: Methods for preparation," *ELs*, March 2002.

[12] S. L. Gerson and M. B. Keagle, *The Principles of Clinical Cytogenetics.* Humana Press, 2nd ed., 2005.

[13] A. Griffiths J.F. et al., *An Introduction to Genetic Analysis.* W.H. Freeman and Company, tenth ed., December 2010.

[14] W. Qiang et al., *Microscope Image Processing.* Elsevier Academic Press, 2008.

[15] M. Moradi et al., "Automatic locating the centromere on human chromosome pictures," in *16th IEEE Symposium on Computer-Based Medical Systems*, pp. 56 – 61, June 2003.

[16] J. Graham et al., "Automatic karyotype analysis," *Chromosome Analysis Protocols*, vol. 29, pp. 141–185, 1994.

[17] S. Gagula Palalic and C. Mehmet, "Extracting gray level profiles of human chromosomes by curve fitting," *Southeast Europe Journal of Soft Computing*, vol. 01, no. 02, pp. 66 – 71, 2012.

[18] N. Otsu, "A threshold selection method from gray-level histograms," *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 9, no. 1, pp. 62–66, 1979.

[19] M. Popescu et al., "Automatic karyotyping of metaphase cells with overlapping chromosomes," *Computers in Biology and Medicine*, vol. 29, pp. 61–82(22), January 1999.

[20] G. Wolf et al., "A pc-based program for evaluation of comparative genomic hybridization (cgh) experiments." http://amba.charite.de/cgh/publ/01/publ01b.html.

[21] J. Kittler and J. Illingworth, "Minimum error thresholding," *Pattern Recognition*, vol. 19, no. 01, pp. 41 – 47, 1986.

[22] D. Somasundaram and V. Kumar, “Straightening of highly curved human chromosome for cytogenetic analysis,” *Measurement*, vol. 47, no. 0, pp. 880 – 892, 2014.

[23] L. Ji, “Fully automatic chromosome segmentation,” *Cytometry*, vol. 17, pp. 196–208, 1994.

[24] V. Gajendran and J. Rodriguez, “Chromosome counting via digital image analysis,” in *International Conference on Image Processing(ICIP)*, pp. 24–27, October 2004.

[25] J. Canny, “A computational approach to edge detection,” *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, vol. 8, pp. 679 – 698, November 1986.

[26] X. Wang et al, “Automated identification of analyzable metaphase chromosomes depicted on microscopic digital images,” *Journal of Biomedical Informatics*, vol. 41, pp. 264–271, 2008.

[27] X. Wang et al., “Automated classification of metaphase chromosomes: Optimization of an adaptive computerized scheme,” *Journal of Biomedical Informatics*, vol. 42, pp. 22 – 31, February 2009.

[28] G. Enrico et al., “Automatic segmentation of chromosomes in q-band images,” in *Proceedings of the 29th Annual International Conference of the IEEE EMBS*, pp. 23–26, August 2007.

[29] M. Kass et al., “Snakes: Active contour models,” *International Journal of Computer Vision*, vol. 1, pp. 321–331, January 1988.

[30] C. Xu and J. L. Prince, “Snakes, shapes, and gradient vector flow,” *IEEE Transaction on Image Processing*, vol. 7, no. 3, pp. 359 – 369, 1998.

[31] P. Britto and G. Ravindran, “Novel findings in chromosome image segmentation using discrete cosine transform based gradient vector flow active contours,” *Information Technology Journal*, vol. 6, no. 1, pp. 1–7, 2007.

[32] C. Li et al, “Segmentation of edge preserving gradient vector flow: An approach towards automatically initializing and splitting of snakes,” in *Proceedings of*

IEEE Computer Society Conference on Computer Vision and Pattern Recognition, vol. 01, pp. 162–167, June 2005.

[33] H. Blum, "A transformation for extracting new descriptors of shape," in *Models for the Perception of Speech and Visual Form*, pp. 362 – 380, MIT Press, 1967.

[34] R. Stanley et al., "Centromere attribute integration based chromosome polarity assignment," in *Proceedings: a conference of the American Medical Informatics Association*, pp. 284–288, 1996.

[35] B. K. Jang and T. C. Roland, "Analysis of thinning algorithms using mathematical morphology," *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, vol. 12, pp. 541 – 551, March 1990.

[36] X. Wang et al, "A rule-based computer scheme for centromere identification and polarity assignment of metaphase chromosomes," *Computer Methods and Programs in Bio Medicine*, vol. 89, pp. 33–42, 2008.

[37] R. M. Mohammad, "Accurate localization of chromosome centromere based on concave points," *Journal of Medical Signals and Sensors*, vol. 02, no. 02, pp. 88 – 94, 2012.

[38] J. Piper and E. Granum, "On fully automatic feature measurement for banded chromosome classification," *Cytometry*, vol. 10, pp. 242–255, 1989.

[39] J. H. Kao et al, "Chromosome classification based on the band profile similarity along approximate medial axis," *The Journal of Pattern Recognition Society*, vol. 41, pp. 77–89, 2008.

[40] G. Ritter and G. Schreib, "Using dominant points and variants for profile extraction from chromosomes," *Pattern Recognition Journal*, pp. 923–938, April 2001.

[41] A. Subasinghe A. et al., "An accurate image processing algorithm for detecting fish probe locations relative to chromosome landmarks on dapi stained metaphase chromosome images," in *Seventh Canadian Conference on Computer and Robot Vision (CRV)*, pp. 223 – 230, May 2010.

[42] S. Jahani and S. K. Satarehdan, "Centromere and length detection in artificially straightened highly curved human chromosomes," in *International Journal of Biological Engineering*, vol. 02, pp. 56–61, 2013.

[43] S. Jahani and S. K. Satarehdan, "A novel algorithm for straightening highly curved images of human chromosome," in *Pattern Recognition Letters* (09, ed.), vol. 29, pp. 1208 – 1217–61, 2008.

[44] P. Mousavi and R. Ward, "Feature analysis and centromere segmentation of human chromosome images using an iterative fuzzy algorithm," *IEEE Transactions on Biomedical Engineering*, vol. 49, pp. 363 – 371, April 2002.

[45] E. R. Faria et al., *Segmentation and Centromere Locating Methods Applied to Fish Chromosomes Images*. Springer-Verlag Berlin Heidelberg, 2005.

[46] T. Kobayashi et al., "Content and classification based ranking algorithm for metaphase chromosome images," in *IEEE Conference on Multimedia Imaging*, 2004.

[47] Mathworks, "www.mathworks.com/access/helpdesk/help/toolbox/images/." web site.

[48] F. Leymarie and M. Levine, "Simulating the grassfire transform using an active contour model," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 14, pp. 56 – 75, Jan 1992.

[49] X. Bai et al., "Skeleton pruning by contour partitioning with discrete curve evolution," *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, vol. 29, pp. 449 – 462, March 2007.

[50] L. J. Latecki and R. Lakämper, "Contour-based shape similarity," in *Visual Information and Information Systems*, vol. 1614/1999 of *Lecture Notes in Computer Science*, p. 657, Springer-Verlag Berlin Heidelberg, January 1999.

[51] C. Xu and B. Kuipers, "Object detection using principal contour fragments," in *Canadian Conference on Computer and Robot Vision (CRV)*, pp. 363–370, May 2011.

[52] L. J. Latecki and R. Lakämper, "Polygon evolution by vertex deletion," in *Proceedings of the Second International Conference on Scale-Space Theories in Computer Vision*, pp. 398 – 409, Springer-Verlag London, UK, 1999.

[53] O. Golubitsky and S. M. Watt, "Online stroke modeling for handwriting recognition," *18th Annual International Conference on Computer Science and Software Engineering (CASCON 2008)*, pp. 72 – 80, 2008.

[54] V. Mazalov and S. M. Watt, "Digital ink compression via functional approximation," *Proc. 12th International Conference on Frontiers in Handwriting Recognition (ICFHR 2010)*, pp. 688 – 694, 2010.

[55] "Maplesoft, a division of waterloo maple inc." http://www.maplesoft.com/, 2014.

[56] "Wolfram." http://www.wolfram.com/mathematica/, 2014.

[57] M. F. Valstar and M. Pantic, "Combined support vector machines and hidden markov models for modeling facial action temporal dynamics," *HumanComputer Interaction*, pp. 118 – 127, 2007.

[58] M. Michael I. et al., "Bayesian construction of geometrically based cortical thickness metrics," *NeuroImage*, vol. 12, no. 6, pp. 676 – 687, 2000.

[59] H. Haidar and J. Soul, "Measurement of cortical thickness in 3d brain mri data:validation of the laplacian method.," *NeuroImage*, vol. 16, pp. 146 – 153, 2006.

[60] J. Stephen E. et al., "Three-dimensional mapping of cortical thickness using laplaces equation," *Human Brain Mapping*, vol. 11, pp. 12–32, 2000.

[61] J. Piper and J. Sprey, "Adaptive classifier for dicentric chromosomes," in *Journal of Radiation Research : 33 Suppliment*, vol. 2, pp. 159 – 170, Mar 1992.

[62] B. R. et al, "Radiation dosimetry by automatic image analysis of dicentric chromosomes," in *Mutation Research/Environmental Mutagenesis and Related Subjects*, vol. 253, pp. 223 – 235, 1991.

[63] S. Biswas and B. C. Lovell, *Chapter 9: Snakes and Active Contours - Bzier and Splines in Image Processing and Machine Vision.* Springer-Verlag Berlin Heidelberg.

[64] P. Britto and G. Ravindran, "Chromosome segmentation and investigations using generalized gradient vector flow active contours." Online Journal of Health and Allied Sciences http://www.ojhas.org/issue14/2005-2-3.htm, 2005.

[65] C. Xu and J. L. Prince, *Handbook of Medical Imaging: Processing and Analysis.* Academic Press, 2000.

[66] L. D. Cohen, "On active contour models and balloons," in *CVGIP: Image Understanding archive*, vol. 53, pp. 211 – 218, Academic Press, Inc, 1991.

[67] L. D. Cohen and I. Cohen, "Finite-element methods for active contour models and balloons for 2-d and 3-d images," *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, vol. 15, no. 11, pp. 1131 – 1147, 1993.

[68] L. Latecki and R. Lakämper, "http://knight.cis.temple.edu/ shape/shape/index.html." web site.

[69] L. J. Latecki et al., "Shape descriptors for non-rigid shapes with a single closed contour," *IEEE conference on Computer Vision and Pattern Recognition (CVPR)*, vol. 1, pp. 424–429, 2000.

[70] L. J. Latecki et al., "Continuity of discrete curve evolution," *Journal of Electronic Imaging*, vol. 09, pp. 317 – 326, July 2000.

[71] K. Siddiqi and A. Shokoufandeh, "http://www.cim.mcgill.ca/ shape/." web site.

[72] A. Madureura et al., *Computational Intelligence and Decision Making - Trends and Applications*, vol. 61. Springer-Verlag Berlin Heidelberg, 2013.

[73] L. Wang, *Support Vector Machines: Theory and Applications*, vol. 177 of *Studies in Fuzziness and Soft Computing.* Springer-Verlag Berlin Heidelberg, 2005.

[74] N. Deng et al., *Support Vector Machines: Optimization Based Theory, Algorithms, and Extensions.* Data Mining and Knowledge Discovery Series, Chapman & Hall/CRC Press, December 2012.

[75] H. W. Kuhn and A. W. Tucker, "Nonlinear programming," *Proceedings of 2nd Berkeley Symposium*, p. 481 492, 1951.

[76] T. Fletcher, *Support Vector Machines Explained.* 2008.

[77] T. Cover, “Geometrical and statistical properties of systems of linear inequalities with applications in pattern recognition,” *Electronic Computers, IEEE Transactions on*, vol. EC-14, pp. 326 – 334, June 1965.

[78] C. Campbell and Y. Ying, “Learning with support vector machines,” *Synthesis Lectures on Artificial Intelligence and Machine Learning*, vol. 05, no. 01, pp. 01 – 95, 2011.

Curriculum Vitae

Name:	
Post-Secondary Education and Degrees:	2008-2010 MESc.(Eng) Electrical & Computer Engineering Western University, Canada.
	2002-2006 B.Sc.(Eng) Electrical Engineering University of Moratuwa, Sri Lanka.
Related Work Experience:	Teaching Assistant The University of Western Ontario 2008 - 2014

Publications:

1. US PATENT. ”Centromere detector and method for determining radiation exposure from chromosome abnormalities.”. United States 8,605,981. PCT No.:PCT/US2011/059257, 2014.

2. Peter K. Rogan, Yanxin Li, Asanka Wickramasinghe, Akila Subasinghe, Natasha Caminsky, Wahab Khan, Jagath Samarabandu, Joan H. Knoll, Ruth Wilkins, and Farrah Flegal.”Automating dicentric chromosome detection from cytogenetic biodosimetry data”, Journal of Radiation Protection Dosimetry, April - 2014.

3. Akila Subasinghe A. et al. ”Intensity integrated Laplacian-based thickness measurement for detecting human metaphase chromosome centromere location”. In IEEE Transactions on Biomedical Engineering (TBME), volume 60, pages 2005 2013, July 2013.

4. W. A. Khan, R. A. Chisholm, S. M. Taddayon, A. Subasinghe, J. Samarabandu, L. J. Johnston, P. R. Norton, P. K. Rogan, J. H. M. Knoll. ”Relating centromeric topography in fixed human chromosomes to a-satellite DNA and CENP-B distribution”, Cytogenetics and Genome Research, 2013.

5. Akila Subasinghe A. et al. ”Intensity integrated Laplacian algorithm for human metaphase chromosome centromere detection”. In Electrical Computer Engineering (CCECE), 2012 25th IEEE Canadian Conference on, May 2012.

6. Rajeev Ranjan, Akila Subasinghe Arachchige, Jagath Samarabandu, Peter K. Rogan and Joan Knoll. ”Automatic Detection of Pale Path and Overlaps in Chromosome Images using Adaptive Search Technique and Re-thresholding”, International Conference on Computer Vision Theory and Applications, 2012.

7. Yanxin Li, Asanka Wikramasinghe, Akila Subasinghe, Jagath Samarabandu, Joan Knoll, Ruth Wilkins, Farrah Flega, and Peter Rogan. ”Towards Large Scale Automated Interpretation of Cytogenetic Biodosimetry Data”, International Conference on Information and Automation for Sustainability, 2012.

8. Akila Subasinghe A, Jagath Samarabandu , Joan Knoll, Wahab Khan and Peter Rogan.”Accurately extracting the centerline from human metaphase chromosomes using image processing”. Canadian Student Conference on Biomedical Computing and Engineering (CSCBCE), 2012.

9. Akila Subasinghe A, Jagath Samarabandu , Joan Knoll and Peter Rogan.”Automated metaphase chromosome centromere refinement using fuzzy inference systems”. Canadian Student Conference on Biomedical Computing and Engineering (CSCBCE), 2012.

10. Akila Subasinghe A. et al. ”An image processing algorithm for accurate extraction of the centerline from human metaphase chromosomes”. In International Conference on Image Processing (ICIP), September 2010.

11. Akila Subasinghe A. et al. ”An accurate image processing algorithm for detecting fish probe locations relative to chromosome landmarks on dapi stained metaphase chromosome images”. In Seventh Canadian Conference on Computer and Robot Vision (CRV), May 2010.

12. Thrishantha Nanayakkara, Lasitha Piyathilaka, Akila Subasingha., ”Mechatronics in Landmine Detection and Removal”, Mechtronic Systems Devices, Design, Control, Operation, and Monitoring, Edited by Clarence De Silva, CRC Press, Taylor & Francis, Boca Raton, FL Chapter 28, 2007.

13. Nanayakkara, T. Piyathilaka, J.M.L.C. Siriwardana, A.P. Subasinghe, S.A.A.M., Jamshidi, M., ”Orchestration of Advanced Motor Skills in a Group of Humans through an Elitist Visual Feedback Mechanism for System of Systems Engineering”, 2007. SoSE ’07