

2010

Image Processing Techniques for Detecting Chromosome Abnormalities

Akila Mike Subasinghe Arachchige
Western University

Follow this and additional works at: <https://ir.lib.uwo.ca/digitizedtheses>

Recommended Citation

Subasinghe Arachchige, Akila Mike, "Image Processing Techniques for Detecting Chromosome Abnormalities" (2010). *Digitized Theses*. 4500.
<https://ir.lib.uwo.ca/digitizedtheses/4500>

This Thesis is brought to you for free and open access by the Digitized Special Collections at Scholarship@Western. It has been accepted for inclusion in Digitized Theses by an authorized administrator of Scholarship@Western. For more information, please contact wlsadmin@uwo.ca.

Image Processing Techniques for Detecting Chromosome Abnormalities.

(Spine title: **Img Processing Tech. for Detecting Chromosome
Abnormalities**)

(Thesis format: **Monograph**)

by

Akila Mike Subasinghe Arachchige

~

**Graduate Program
in
Engineering Science
Electrical and Computer Engineering**

**A thesis submitted in partial fulfillment
of the requirements for the degree of
Master of Engineering Science**

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

© Akila S. 2010

Certificate of Examination

THE UNIVERSITY OF WESTERN ONTARIO
FACULTY OF GRADUATE STUDIES
CERTIFICATE OF EXAMINATION

Chief Advisors:

Dr. Jagath Samarabandu

Dr. Peter Rogan

Advisory Committee:

Dr. Joan Knoll

Examining Board:

Dr. Luiz Capretz

Dr. Ken McIsaac

Dr. Lucian Ilie

The thesis by

Akila Mike Subasinghe Arachchige

entitled:

Image Processing Techniques for Detecting Chromosome Abnormalities.

is accepted in partial fulfillment of the
requirements for the degree of
Master of Engineering Science

Date: _____

Chair of Examining Board
Dr. Jayshri Sabarinathan

Abstract

With the increasing use of Fluorescence In Situ Hybridization (FISH) probes as markers for certain genetic sequences, the requirement of a proper image processing framework is becoming a necessity to accurately detect these probe signal locations in relation to the centerline of the chromosome. Automated detection and length measurements based on the centerline relative to the centromere and the telomere coordinates would highly assist in clinical diagnosis of genetic disorders and thus improve its efficiency significantly. Although many image processing techniques have been developed for chromosomal analysis such as "karyotype analysis" to assist in laboratory diagnosis, they fail to provide reliable results in segmenting and extracting the centerline of chromosomes due to the high variability in shape of chromosomes on microscope slides.

In this thesis we propose a hybrid algorithm that utilizes Gradient Vector Flow active contours, Discrete Curve Evolution based skeleton pruning and morphological thinning to provide a robust and accurate centerline of the chromosome, which is then used for the measurement of the FISH probe signals. Then this centerline information is used to detect the centromere location of the chromosome and the probe signal location distances were measured with respect to these landmarks. The ability to accurately detect FISH probe locations with respect to its centerline and other landmarks can provide the cytogeneticists with detailed information that could lead to a faster diagnosis.

Acknowledgements

I would like to thank my main supervisor, Dr. Jagath Samarabandu for all the invaluable guidance, advice and patience lent throughout my masters, both academically and personally. Dr. Samarabandu 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. Jagath Samarabandu, Dr. Olga Veksler, Dr. Yuri Boykov and Dr. John Barron for their innovative and attractive ways of teaching and motivating my work.

Heather Tarnowski & Wahab Khan helped me with acquiring DAPI images for my research and I am greatly thankful for their help. I would like to thank the IRIS research group colleagues (Eranga Ukwatta, Mehrnaz Zougi, Edwin Noroshan & Tharindu De Silva) for the interesting discussions which in many ways had a positive influence on my research.

I would also 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, support and character lent and shown from 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. Last, but 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 terms was actively involved in this thesis.

Akila Mike,
June - 2010,
London ON.

Table of Contents

Certificate of Examination	ii
Abstract	iii
Acknowledgements	iv
List of tables	vii
List of figures	viii
Acronyms	xi
1 Introduction	1
1.1 Introduction to human chromosomes	3
1.1.1 Human chromosome structure	3
1.2 Contributions	7
1.3 Thesis organization	9
2 Methods	10
2.1 The experimental setup	10
2.2 Literature review	12
2.2.1 Segmentation methods	12
2.2.2 Centerline detection	14
2.3 GVF snakes	17
2.3.1 Parametric snakes or active contours	17
2.3.2 Gradient vector flow as an external energy	21
2.3.3 GVF snakes vs DT snakes	22
2.4 Discrete curve evolution (DCE)	24
2.4.1 Definitions	25
2.4.2 The relevance function	27
2.4.3 Advantages & disadvantages of DCE	29
2.5 Cubic spline interpolation	31
2.5.1 Definitions	32
2.5.2 Spline boundary conditions	34

Table of Contents

3	Proposed algorithm	36
3.1	Pre-processing and segmentation	37
3.2	Centerline detection	47
3.3	Chromosome end point detection	57
3.4	Centromere identification and polarity assignment	59
3.4.1	The centromere confidence measure	62
3.5	FISH probe projection	66
3.6	Fractional ratio measurement	71
4	Results and discussion	75
4.1	Preliminary testing	75
4.2	Quantitative analysis	78
4.2.1	Error metric	81
4.2.2	Test results	83
4.2.3	Variabilities or dependencies	85
5	Conclusion and future work	89
5.1	Summary and conclusion	89
5.2	Future work	90
 Appendices		
	Curriculum Vitae	97

List of Tables

1.1	Typical range of centromere index (CI) for each human chromosome [1]	8
3.1	The 4 connected removal process used to obtain the contour of the selected object. Here, the middle pixel of 3.1(a) is flipped from '1' to a '0' as in 3.1(b) based on the 4-connected neighbors	45
3.2	A 7x7 representation of the original 20x20 template used for end point correction, where the coefficients were set as '0' - ignored, '+' as +1 and '-' as -1.	58
4.1	Expert scoring results for the centromere location detection accuracy in two data sets. Here, the term 'neighboring' is used when the detected centromere location is within 1 chromosomal band distance from the actual centromere location. The sensitivity is calculated by considering both 'accurate' and 'neighboring' as acceptable results (true positives)	76
4.2	Comparative scoring of DCE vs thinning algorithms - where each value gives the number of chromosomes for which, the centerline was better represented by the corresponding algorithm	77
4.3	Matric results (MAD and MAXD) mean values and standard deviation for the complete analyzed data set.	84
4.4	Matric results (MAD and MAXD) mean values and standard deviation for each chromosome group. The number (within brackets) after each chromosome group name specifies the number of chromosomes present in the data set from that particular group.	84
4.5	Matric results (MAD and MAXD) mean values and standard deviation for 16 chromosomes from patient No.1, 21 chromosomes from patient No.2 and 10 chromosomes from patient No.3.	85
4.6	Matric results (MAD and MAXD) mean values and standard deviation for 10 chromosomes from slide No.1 and 11 chromosomes from slide No.2 from the same patient.	86

List of Figures

1.1	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).	2
1.2	An illustration structure of a chromosome in context of the cell image (Source- http://commons.wikimedia.org/wiki/File:Chromosome.gif). .	4
1.3	The structural components of metaphase chromosomes of all three categories : the metacentric, sub-metacentric and acrocentric	6
1.4	The lengths used for calculating the centromere index of a given chromosome.	7
2.1	The experimental setup including the anatomy of a fluorescence microscope.	11
2.2	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) [2].	23
2.3	Comparison between Distance Potential (DT Based) vector field (left) and the GVF vector field (right) [2].	24
2.4	above: Steps of the DCE process of shape simplification, below: the same steps when initiated by adding boundary noise to the same image [3]. (Reproduced with permission from Prof. Longin Jan Latecki)	27
2.5	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'.	27
2.6	Shape variations of polygon partitions and the effects of turn angle and arc length to the relevance measure.	29
2.7	Spline curve segment arrangement	32
3.1	The flow chart of the Pre-processing stage algorithm.	38
3.2	A digital cell image with two FISH probes located on chromosome 15 and its homologue. This also depicts the difficulty (image processing wise) in locating these small probe signals in a chromosome cell image.	39
3.3	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).	41

3.4	A comparison of the intensity histograms of the chromosome images with from top to bottom: Original cropped image, contrast enhanced image and the histogram equalized image.	42
3.5	Examples of segmentation of two extracted image windows using Otsu method where the boundary discontinuities of the Otsu result are evident.	44
3.6	Application of different image filters (mean, Gaussian and median) and their corresponding edge maps. All the filters were applied to a neighborhood of 5×5 region.	46
3.7	Two chromosomes and their skeletonization [4] and morphological thinning results [5] showing some resulting spurious branches. These operations were performed on the binary object obtained through Otsu's method.	49
3.8	Comparison between standard skeleton with DCE based solutions where the figure 3.8(b) is the skeletonization result. The marked points figure 3.8(c) & figure 3.8(d) depict the end polygon vertices of the DCE process with different end conditions.	53
3.9	The 3×3 neighborhood setup for the thinning process where p is the pixel of interest and $X_9 = X_1$	54
3.10	An example rectification achieved through the Hit & Miss algorithm. The location of interest (for rectification) is circled for clarity.	55
3.11	The masks used for the Hit & Miss process where '-1' as -1, '+1' as 1 and '0' as ignored.	55
3.12	The trellis (in yellow) and the superimposed GVF contour (in blue) used for the centromere identification.	61
3.13	The width profile (figure 3.13(b)) and the intensity profile (figure 3.13(c)) of a straight chromosome which is shown in figure 3.13(a). The x axis in figure 3.13(b) & figure 3.13(c) is the sample data points of the centerline (from bottom to top of the chromosome) while the y axis are respectively the width and the image intensity values at that point.	64
3.14	The width profile (figure 3.14(b)) and the intensity profile (figure 3.14(c)) of a 'bent' chromosome which is shown in figure 3.14(a). The x axis in figure 3.14(b) & figure 3.14(c) is the sample data points of the centerline (from top to bottom of the chromosome) while the y axis are respectively the width and the image intensity values at that point. The points which corresponds to the minimum width profile value is marked in figure 3.14(b).	65
3.15	The tangent based method setup.	66
3.16	A possible fail scenario for the geometric iteration method. In this depicted situation, the point 'Q' (through Q_1, Q_2, \dots) would not converge to the point closer to the point 'P' which is the desired result.	68

List of Figures

3.17	An instance where the point projection method could fail due to high bends on the chromosome, where the objective of the algorithm is to find a point closest to the 'yellow circle' on the curve. The 'blue points' shows projected points obtained at each iteration while the 'magenta' point depicting the final converged result.	69
3.18	A limitation of the nearest neighbor approach where S_{n-1} , S_n & S_{n+1} are three points on the centerline with same Euclidian distance (d) from the point (FISH probe) 'P'.	70
3.19	The flow chart of the point projection method used in the research. .	72
3.20	The diagram depicting the known length measurements and positions which were detected through our proposed algorithm.	73
4.1	End point corrected centerline (black & white line) results of representative bent chromosomes using the proposed approach.	78
4.2	Some of the chromosome centromere detection results with their respective chromosome groups.	79
4.3	Point projection results - (yellow square - test point : blue circle - projected point)	80
4.4	The configuration used by Ladak to analyze accuracy of a 2D prostate boundary segmentation where the black contour and the red contour represents the gold standard and the actual segmentation respectively [6]. The measurement value $d(\theta)$ was carried out along the line segments (C - A1 etc.) based on the radial angle value θ (between C-C' and C-A1)	81
4.5	The configuration used in our analysis where the black and the red curve segments represent the gold standard and the obtained centerline respectively. The measurement value $d(\theta)$ was obtained for an example, between points P and R, along the line segments (C - C' etc.)	82
4.6	MAD and MAXD (centerline) metric mean and std. deviation values of chromosome groups A to E.	87
4.7	Comparison of MAD and MAXD (centerline) metric mean values of chromosome groups A to E obtained through the proposed (DCE based) and thinning methods.	88

Acronyms

CI	<i>Centromere Index</i>
DAPI	<i>4',6-Diamidino-2-Phenylindole</i>
DCE	<i>Discrete Curve Evolution</i>
DT	<i>Distance Transform</i>
FITC	<i>Fluorescein Isothiocyanate</i>
GVF	<i>Gradient Vector Flow</i>
G-banded	<i>Geimsa Banded</i>
MAT	<i>Medial Axis Transform</i>
NN	<i>Nearest Neighbor</i>
MAD	<i>Mean Absolute Difference</i>
MADX	<i>Maximum Absolute Difference</i>
PC	<i>Parametric Curve</i>

Chapter 1 Introduction

Many chronic diseases can be traced back to the DNA structure of a patient. Therefore the study of human chromosomes and their structure is of utmost importance in clinical diagnosis. Non radioactive Fluorescence In Situ Hybridization (FISH) has 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, which could be used to detect certain chromosomal abnormalities [7]. Dr. Rogan and Dr. Knoll have developed a labeling technology called 'single-copy DNA probe technology' which can be used to diagnose genetic disorders such as leukemia [8]. These DNA probes have very high resolution (small length) and therefore can be tailored to label any part of the human genome. Therefore, by varying these remarkably small probes, cytogeneticists can study many chromosome abnormalities such as translocation, deletion etc. at a very fine resolution. Also, this high resolution can lead to finding new dependencies between genetic diseases and chromosome aberrations. An automated image processing application can be very useful in analyzing these dependencies and variations. Thus both diagnosis and research time can be cut down by a drastic margin.

Its important to first understand the characteristics of these probes which we intend to analyze. FISH uses fluorescence DNA probes to detect chromosome sequence rearrangements in genetic diseases. Karyotype analysis is one of the main research areas in image processing which aims at producing annotated karyograms with the least user involvement (refer figure 1.1). Methods available for karyotyping or other chromosome analysis are mainly limited by the shape variability caused by non-rigid nature of the chromosome structure. Therefore, the effectiveness of these image processing techniques are limited by the inability to provide proper results irrespective of the shape of the chromosome [9]. Proper segmentation and extraction of the center line of the chromosome plays a vital role in many of the available karyotype analysis methods [10],[11]. In this research, our image processing techniques are applied to DAPI (4',6-Diamidino-2-Phenylindole) stained chromosome images in contrast to Geimsa stained images used in many karyotype analysis methods in literature.

Karyotype analysis is used in this paper merely for comparison of methodologies used for information extraction. This proposed algorithm can be readily adopted in any type of analysis which needs similar information. In this research, we have identified the following as the most important steps for analyzing FISH probe signals with respect to the chromosome structure [11],[12],[13],

- An accurate segmentation of the chromosome.

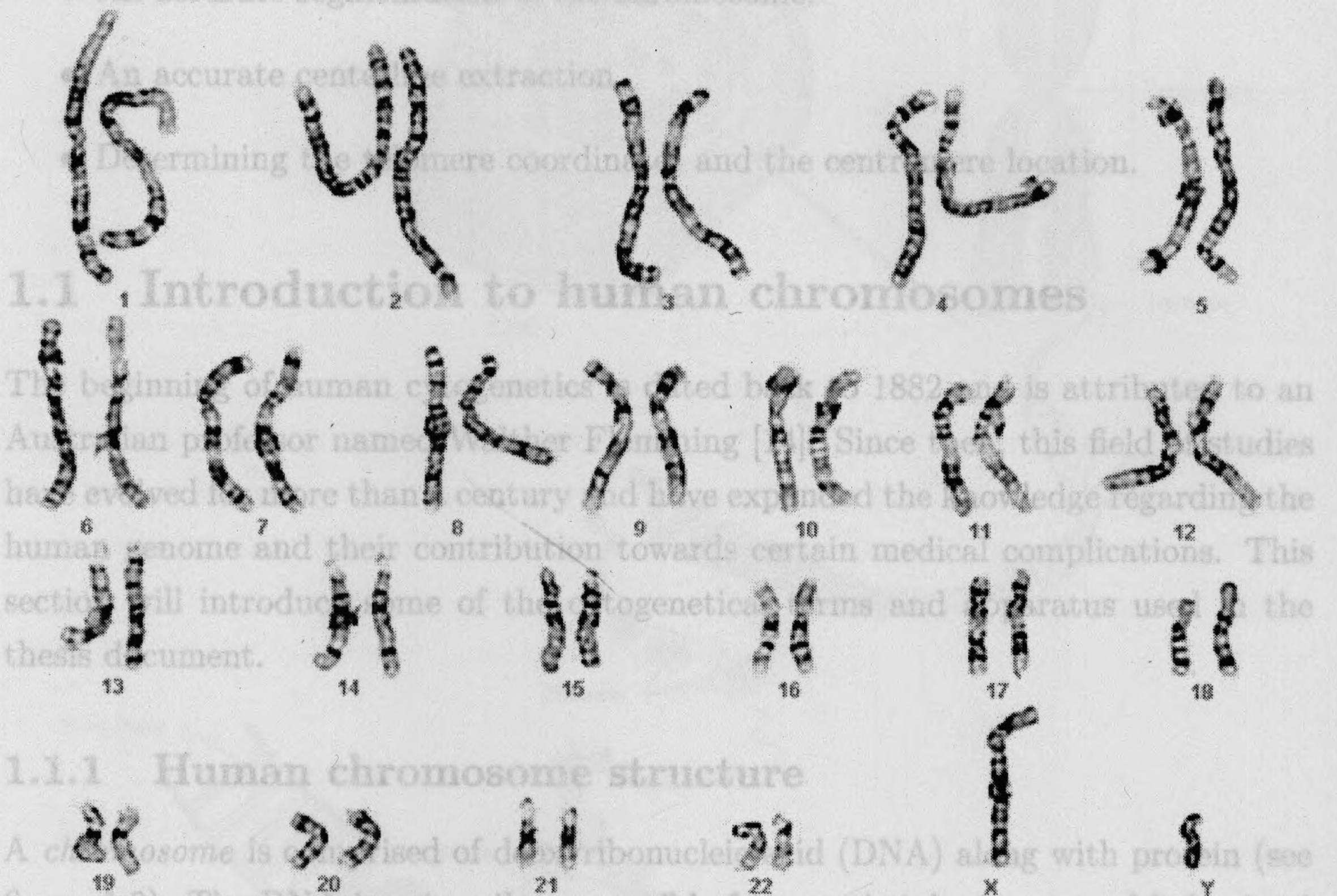


Figure 1.1: A karyotype analysis end result with all 46 chromosomes organized according to their chromosome groups and types (Reproduced with permission from Dr. Joan Knoll and Dr. Peter Rogan).

Karyotype analysis is used in this paper merely for comparison of methodologies used for information extraction. This proposed algorithm can be readily adopted in any type of analysis which needs similar information. In this research, we have identified the following as the most important steps for analyzing FISH probe signals with respect to the chromosome structure [11],[12],[13],

- An accurate segmentation of the chromosome.
- An accurate centerline extraction.
- Determining the telomere coordinates and the centromere location.

1.1 Introduction to human chromosomes

The beginning of human cytogenetics is dated back to 1882 and is attributed to an Australian professor named Walther Flemming [14]. Since then, this field of studies have evolved for more than a century and have expanded the knowledge regarding the human genome and their contribution towards certain medical complications. This section will introduce some of the cytogenetical terms and apparatus used in the thesis document.

1.1.1 Human chromosome structure

A *chromosome* is comprised of deoxyribonucleic acid (DNA) along with protein (see figure 1.2). The DNA is primarily responsible for genetic inheritance and behavioral patterns of a human being. A chromosome consists of two *sister chromatids* which are contracted together forming a double helical shape. A healthy person would have 46 chromosomes as 44 *autosomes* (i.e.- not sex chromosomes) and two sex chromosomes X and Y. The 44 autosome are numbered from 1-22 in the descending order of the length, size and the centromere position of each of these pairs [14]. Presence of two XX chromosomes normally represents a female while XY would normally specify a male.

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

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

During this research we have observed the centromere and telomere regions.

1.1.1.1 The centromere

The centromere of a chromosome is the region where the two sister chromatids are joined and twisted together during *mitosis* (cell division) [15]. The centromere is an essential part of the chromosome that aids splitting of sister chromatids in the cell division process. In many groups of chromosomes, this region is observed as a clear constriction in the width profile of the chromosome. Human chromosomes can be grouped into 3 categories based on the location of the centromere with respect to its ends.

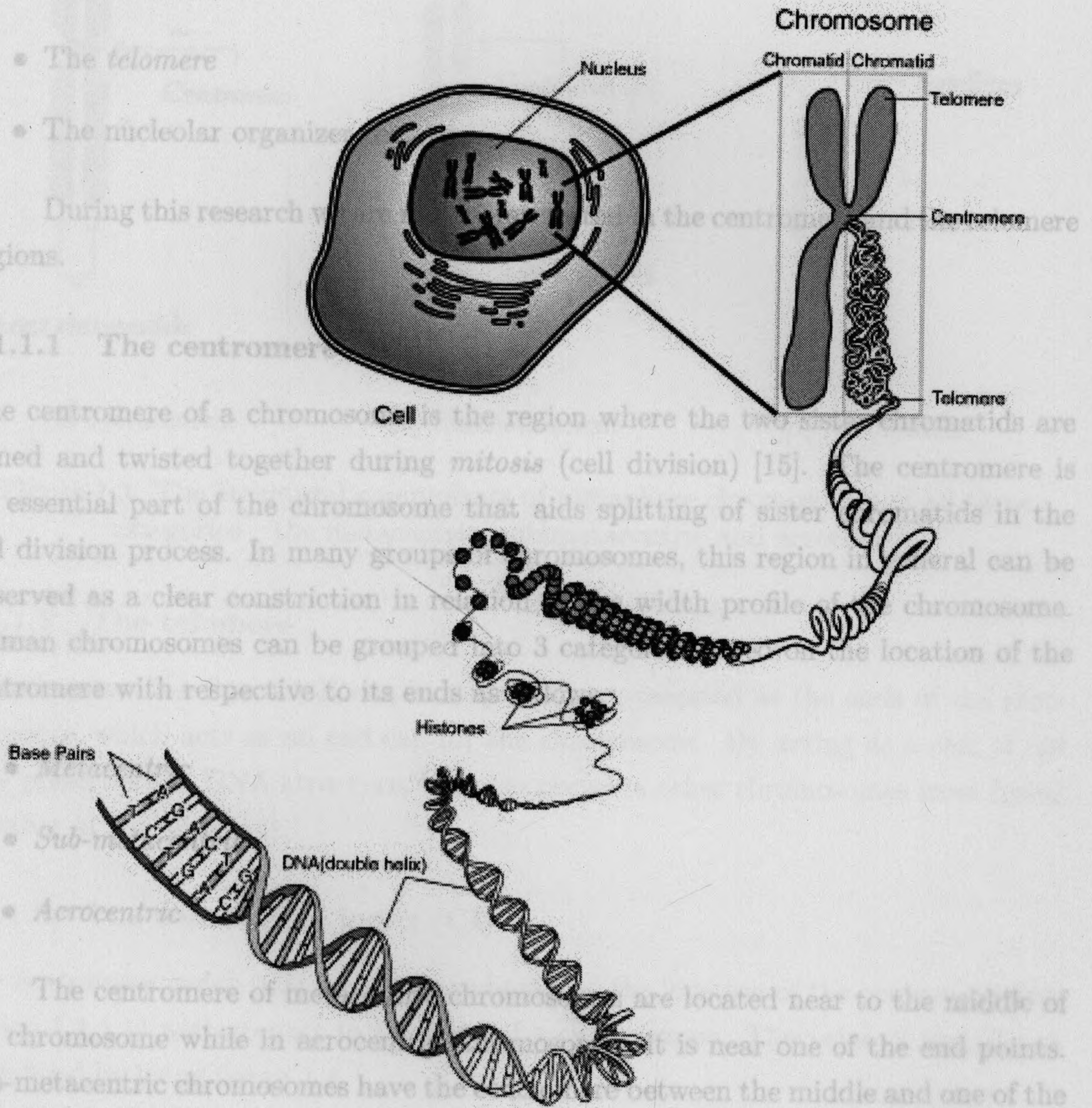


Figure 1.2: An illustration structure of a chromosome in context of the cell image (Source- <http://commons.wikimedia.org/wiki/File:Chromosome.gif>).

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

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

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

1.1.1.1 The centromere

The centromere of a chromosome is the region where the two sister chromatids are joined and twisted together during *mitosis* (cell division) [15]. The centromere is an essential part of the chromosome that aids splitting of sister chromatids in the cell division process. 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. Human chromosomes can be grouped into 3 categories based on the location of the centromere with respect to its ends as follows,

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

The centromere of metacentric chromosomes are located near to 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 1.3

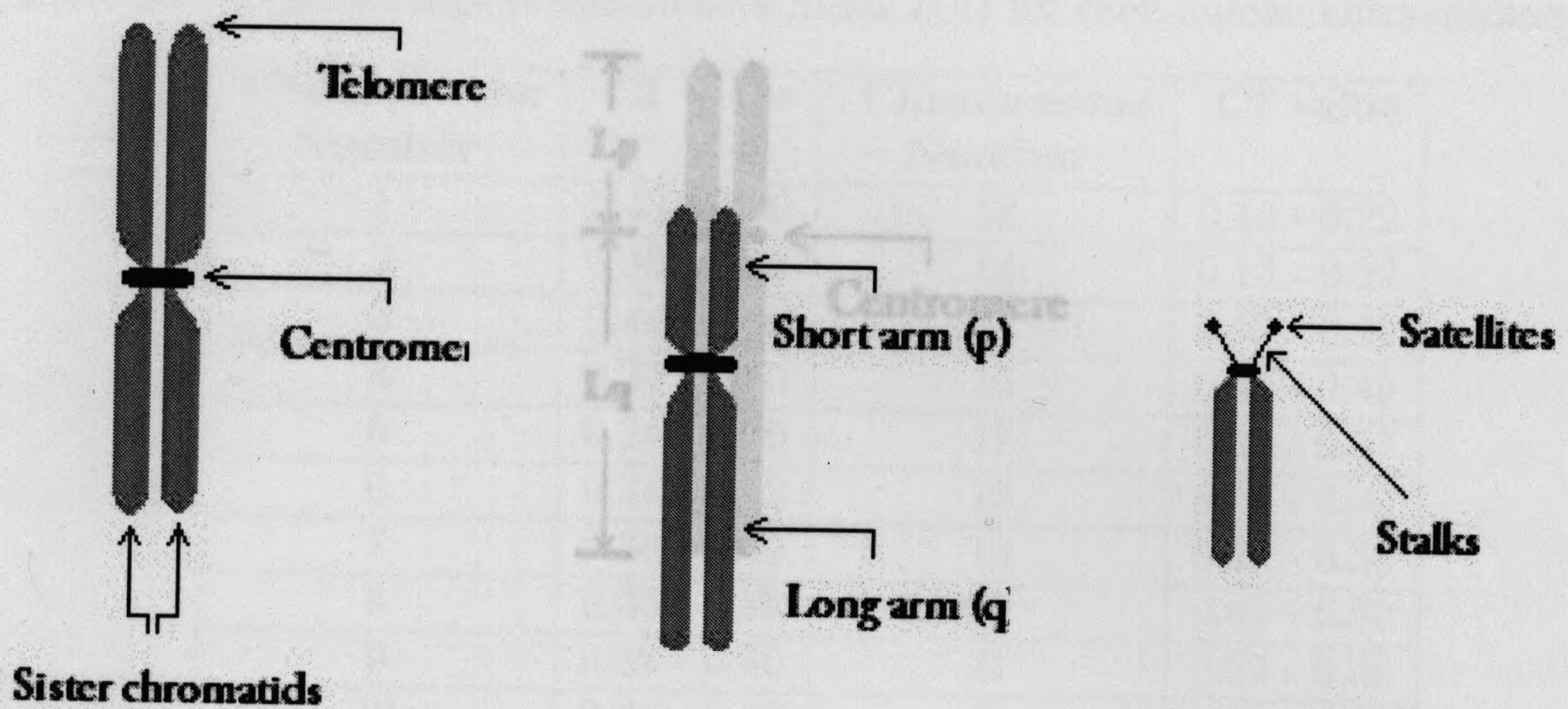


Figure 1.4: The lengths used for calculating the centromere index of a given chromosome are shown in the diagrams. (a) Metacentric (b) Submetacentric (c) Acrocentric

Figure 1.3: The structural components of metaphase chromosomes of all three categories : the metacentric, sub-metacentric and acrocentric

1.1.1.2 The telomere

The telomere (see figure 1.3) can be physically represented as the ends of the chromosomes, which acts as an end cap for the chromosome. By acting as a cap, it not only preserves the DNA structure, but also prevents other chromosomes from fusing together at the ends [14].


1.1.1.3 The centromere index (CI)

The *centromere index* (CI) is a measure based on the location of the centromere (see section 1.1.1.1) with respect to the ends of a chromosome. The value of the CI can be defined by using figure 1.4.

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 can be stated as,

$$CI = \frac{L_p}{L_p + L_q} \quad (1.1)$$

Table 1.1: Typical range of centromere index (CI) for each human chromosome [1]



Chromosome Number	CI	Chromosome Number	CI value
1	0.45	13	0.13 - 0.22
2	0.35	14	0.13 - 0.22
3	0.44	15	0.13 - 0.22
4	0.34	16	0.41 - 0.45
5	0.24	17	0.28 - 0.37
6	0.34	18	0.23 - 0.33
7	0.24	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
12	0.24	Y	0.28 - 0.34

Figure 1.4: The lengths used for calculating the centromere index of a given chromosome.

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 information that can be used to assist chromosome identification and classification. Table 1.1 below provides these values for reference [1].

1.2 Contributions

Our main objective in this thesis is to develop a set of algorithms that will help locate and quantitatively analyze the locations of fluorescent probes. As mentioned above, this deals mainly with three image processing steps:

1. An accurate segmentation of the chromosome
2. An accurate centerline extraction
3. Determining the telomere coordinates and the centromere location

This dissertation presents a novel hybrid image processing algorithm that can be utilized to obtain fractional ratio measurements of FISH probes in DAPI stained metaphase and pro-metaphase chromosomal images.

Table 1.1: Typical range of centromere index (CI) for each human chromosome [1]

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

The following are the main contributions of this algorithm when compared with the state of the art,

1. Development of a semi-automated algorithm for FISH probe detection and fractional ratio measurements in DAPI stained chromosomes.
2. A methodology to obtain an accurate centerline of a chromosome, which is more robust to chromosome shape deformations and boundary noise.
3. Defining measurement ratios that can be utilized in detecting abnormalities of the chromosome structure (chromosome translocation, deletion etc)
4. A measure, termed the 'centromere confidence' was introduced which could represent a rough approximation to the accuracy or confidence in the centromere detection process.
5. A hybrid (intensity and template matching) approach for detecting the telomeric regions of a chromosome
6. An image processing application which directly provides meaningful information efficiently to cytogeneticists. Therefore, could drastically speed up the diagnosis process by identifying chromosome abnormalities such as translocation, deletion etc...

7. An image processing application to measure FISH probe location (ratio measurements) on chromosomes was presented. Therefore, the presentation of this information to cytogeneticists, could drastically speed up the diagnosis process (by identifying chromosome abnormalities such as translocation, deletion etc.).

Some of the work presented in this thesis is accepted for publication at the following conferences,

1. 'International Conference on Image Processing' (ICIP 2010) , Hong Kong, 26 - 29 September 2010 - under the title 'An Image Processing Algorithm for Accurate Extraction of the Centerline from Human Metaphase Chromosomes'.
2. 'Seventh Canadian Conference on Computer and Robot Vision' (CRV 2010) , Ottawa, Ontario, 31 May - 2 June 2010 - under the title 'An Accurate Image Processing Algorithm for Detecting FISH Probe Locations Relative to Chromosome Landmarks on DAPI Stained Metaphase Chromosome Images'.
3. 'International Biodosimetry workshop' Ottawa, Ontario, May 19th 2010 - under the title 'Towards automated single copy FISH probe and centromere identification on metaphase chromosomes'.

1.3 Thesis organization

In this chapter we have discussed the problem domain addressed by the proposed algorithm.

Chapter 2 provides existing solutions in literature, along with their strengths and weaknesses. Also it provides a theoretical analysis of all major algorithms and methods used for our research. The proposed algorithm is discussed in Chapter 3 by dividing it into several functional stages.

Chapter 4 presents the results of applying our proposed algorithm to real DAPI stained metaphase chromosomes. The results are then compared with morphological thinning to obtain an estimation of performance and accuracy. This chapter is followed by conclusive remarks and potential future work in chapter 5.

Chapter 2 Methods

The main objective of this chapter is to provide an overall theoretical view of some of the algorithms utilized for the proposed method and to review some of the existing techniques in literature. First, the experimental setup used for this research will be explained. This will be followed by a detailed literature review and theoretical concepts used for this research. Following are the concepts that will be discussed in this chapter,

- Gradient Vector Flow (GVF) Snakes
- Discrete Curve Evolution
- Cubic Spline curve fitting

2.1 The experimental setup

The experimental setup in this research consisted of an automated fluorescence microscopy system equipped with digital capturing capabilities. The microscopy system was set up by equipping a regular brightfield microscope with an epifluorescence lamp housing, a horizontal attachment for the fluorescent light path, fluorescent filters and fluorescent objective lenses. A detailed illustration of the experimental setup and the anatomy of a fluorescence microscope is provided in figure 2.1. A fluorescent stain is excited by one wavelength of light which then causes the emission of a second wavelength. The final digital images were captured at a magnification of 100X, and were then transferred to a desktop computer for further processing.

From these large set of images, only a handful (less than 5%) would consist of interpretable metaphase chromosomes [16]. Therefore, immediately after storing, these images (from the same slide) were subjected to ranking according to the 'content and classification based ranking' (CCBR) algorithm [16], which stored the ranked images based on the quality of the spread of the chromosome image. This algorithm extracts 17 features (e.g.- area, width, perimeter, overlap ratio of chromosomes etc.)

from all the given chromosome cell images. Then it classifies them into three different groups termed 'nice', 'overlap' and 'spread', based on the lay of chromosomes on the slide. Next, using a virtual query (constructed from the 'nice' group), it ranks all the images starting from the 'nice' images. Here, the 'overlap' images are preferred over 'spread' images in order to ensure the presence of all 46 autosomes in the cell image.

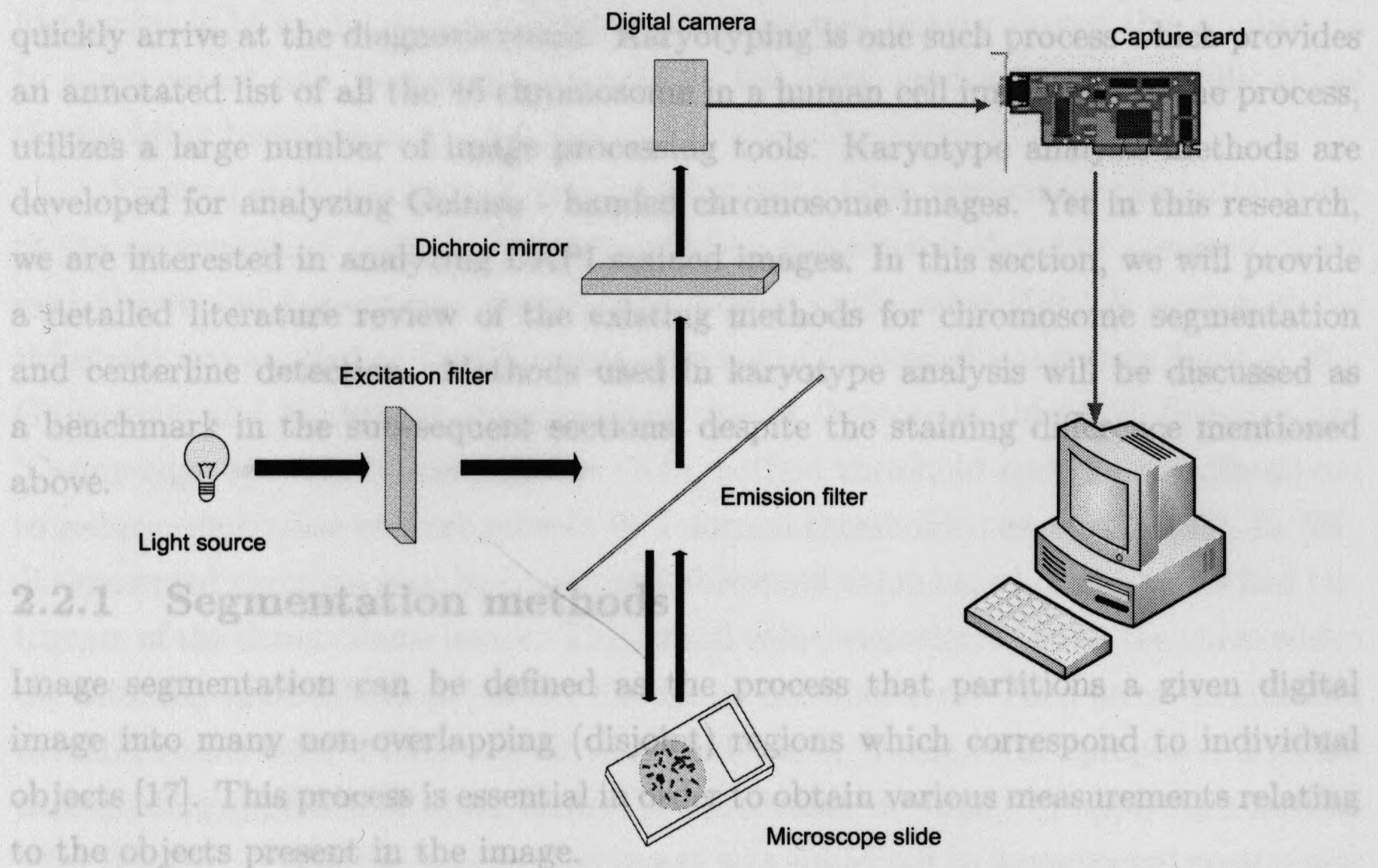


Figure 2.1: The experimental setup including the anatomy of a fluorescence microscope.

In this research, these ranked digital images were acquired through a setup at the 'laboratories of genome bioinformatics and genomic disorders'¹. Algorithms for this research were developed and tested on mainly using MATLAB programming language with the intension of extending to C/C++ programming platforms to increase the efficiency.

1. Prof. Rogan, Schulich School of Medicine and Dentistry, The University of Western Ontario, CANADA

2.2 Literature review

With the invention of new chromosome staining methods, the demand is rising for automated microscopy systems. Yet, the speed of the diagnosis process is highly limited by the time required for a trained cytogeneticist to examine the chromosome images. Therefore, a lot of research has been carried out to assist the cytogeneticist to quickly arrive at the diagnosis result. Karyotyping is one such process which provides an annotated list of all the 46 chromosome in a human cell image and in the process, utilizes a large number of image processing tools. Karyotype analysis methods are developed for analyzing Geimsa - banded chromosome images. Yet in this research, we are interested in analyzing DAPI stained images. In this section, we will provide a detailed literature review of the existing methods for chromosome segmentation and centerline detection. Methods used in karyotype analysis will be discussed as a benchmark in the sub-sequent sections, despite the staining difference mentioned above.

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 [17]. This process is essential in order to obtain various measurements relating to the objects present in the image.

Chromosomes demonstrate high variability in shape (on microscope slides) mainly due to the different stages of the cell cycle, slide preparation and banding patterns among many other small dependencies. This diversity of morphology presents a significant challenge in segmenting as well as in extracting the centerline of a chromosome. Therefore some of the authors have attempted to use manually segmented chromosomes in their research. For an example, Moradi [18],[19] used chromosomes that were manually extracted by an expert.

Various methods are available in literature which attempt to segment chromosomes effectively using various image processing techniques, despite the above challenge.

2.2.1.1 Point processing methods

Most of the existing approaches in literature rely on a form of thresholding, at least as a part of the segmentation process [12]. Thresholding is a simple point processing method. It performs well when subjected to images consisting objects with contrasting pixel intensities with that of the background. In other words, when the image histogram is bi-modal and separable. Chromosomal images possess this quality up to some extent and therefore, thresholding has been used extensively while global thresholding has been the most common.

Popescu et al. performed segmentation using Otsu method for the initial stage of the segmentation (called 'over-segmentation stage') and therefore assigned pixels into object and background regions based on a single value [13]. A similar global threshold was applied to DAPI images (instead of G-banded images) by Wolf in [20]. Gajendran and Rodriguez proposed the use of 'hysteresis thinning' (used in the 'Canny edge operator') based on the Otsu method threshold and thereby attempted to reduce some noise content present in a normal thresholded image [21],[22]. In [23], Ji 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 little higher threshold value. Wang et al. proposed a global thresholding approach in which he first pre-processed the image by applying a median filter [24],[25]. Then, the thresholded image was subjected to 4-connected component labeling to remove isolated noise in the binary image.

Though some of these methods were successful in removing noise from the binary image, they all were prone to lighting effects in the image. Uneven illumination in the image could cause the thresholded objects to be noisy and even discontinuous at some locations. Therefore, attempts were made to segment chromosomes using local or adaptive thresholding and therefore compensating for the lighting variations. In one such attempt, Enrico et al. first divided the image into tessellations of fixed size (manually set) and then thresholded based on the Otsu method [26].

The thresholded image result obtained from any of the above mentioned methods is highly sensitive to quantization errors present in the digital image (see figure 3.5). Also, it is sensitive to intensity fading which is specially found near the boundary regions of chromosomes. Therefore the outline of the segmented object is

often not accurate and doesn't demonstrate the precise shape variations along chromosome boundary. So the pure thresholded image is less suited for further measurements, unless it is refined using another method. Furthermore, due to small variations in illuminance (especially in fluorescence microscopy), the thresholded objects may even contain holes.

2.2.1.2 Other methods

Another segmentation method found in literature is parametric deformable models. Among these, Gradient Vector Flow (GVF) based active contours have been proven to deliver better results, especially in chromosome image segmentation. This deformable model addresses a main limitation in the traditional active contours [27] by drastically improving its capture range [28]. The works of Britto & Ravindran and also Li et al. has reported significant improvements in chromosome segmentation by using the GVF snake model [29],[30]. Yet, this been a parametric active contour, the global minima is not guaranteed unless the control points are initialized in the vicinity of the desired contour. Otherwise, the contour could converge to an unwanted local minima such as a chromosomal band (which has a strong intensity gradient) or even to the contour of another chromosome.

2.2.2 Centerline detection

The centerline of a chromosome is an important feature that in return can be used as reference for numerous chromosome measurements. Therefore, the centerline can be directly used to obtain or sample the following,

- The total length of the chromosome.
- The centromere location and then the centromere index value.
- The coordinates of the telomeric regions of any given chromosome.
- The banding pattern of a chromosome which could be used to identify a chromosome and to classify it accordingly.

We will look at some of the existing methods for finding the centerline of a chromosome in the subsequent sections (section 2.2.2.1 & section 2.2.2.2).

2.2.2.1 Medial axis transform or thinning based methods

Medial Axis Transform (MAT) and morphological thinning are the most commonly adopted methods in finding the centerline of chromosomes. 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 MAT to get 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 [19]. 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.

Thinning is another very common method, which gives less spurious branches compared to skeletonization. The thinned result usually has missing data near the ends of the chromosome [5],[31]. Therefore, it is usually accompanied by an end point extension method in literature. In one such attempt, Wang et al. applied morphological thinning to the segmented binary object and then sampled with a 5-pixel interval. Then these points were interpolated to obtain the chromosome centerline [32],[25]. Gajendran & Rodriguez applied thinning to a median filtered digital image and thereby obtained the centerline [21]. But both the above mentioned approaches are prone to any spurious branches as well as any bifurcations near the telomere locations. Thinning also tends to produce unwanted branches, though not to the same extent as in skeletonization. In digital images where many boundaries are rough due to quantization, despite numerous filtering methods, the spurious branches are eminent. Therefore, in both MAT and thinning methods, the pruning becomes the real limitation of applicability. On top of that, these two methods provide a set of points in space, in contrast to the desired parametric curve that can be effectively and easily used for further calculations.

2.2.2.2 Other centerline finding methods

Methods based on MAT or thinning are subjected to many inherent limitations (see section 2.2.2.1). Therefore a lot of research work has been carried out in order to find suitable methods without using skeletonization or thinning. Jim Piper and Erick Granum proposed a two stage approach in finding the centerline in which they first determined the orientation of the chromosome by calculating the minimum width enclosing rectangle [9]. Then if the chromosome is not highly bent, it was rotated for the orientation to be vertical and mid points of the horizontal chromosome slices were connected together to get the centerline which was then smoothed to get the "poor man's skeleton" (PMS). If the chromosome is bent, they resorted to conventional skeletonization [4] for the centerline. Yet, the problem with this approach is the spurious branches obtained at the conventional skeletonization process, which is depicted in Figure 3.7. Gunter Ritter proposed a method which is based on finding the dominant points of the chromosome [11]. Nevertheless, as reported under results & discussion in [11], reliable results were not obtained when exposed to highly bent chromosomes as well as blurred chromosomes in the above mentioned method. In another approach, chromosomes were sampled into scan lines of different inclinations (0° , 45° , 90° & 135°) and after selecting proper cross-sections, the selected mid points were combined to obtain an approximate centerline [10]. The drawback of this method is that it attempted at getting a polygonal approximation to the centerline instead of the centerline itself. Poor results were obtained when the segmented chromosome boundaries were irregular in shape, which is a commonly encountered situation when handling medical imaging.

2.3 GVF snakes

Gradient vector flow (GVF) snakes is a widely used active contour model in segmentation. It is a well known method formulated in order to have better convergence at boundary concavities. Section 2.3.1 will provide a brief overview of the traditional active contours and will set up the platform for section 2.3.2 which discusses the GVF snake external energy model in detail. Then a brief comparison between the GVF snakes and Distance Transform (DT) based snakes will be stated in section 2.3.3.

2.3.1 Parametric snakes or active contours

Parametric active contours were first introduced by Kass et al. [27] in 1988 and have been applied to many image processing problems throughout the years. 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 [33]. A parametric curve (PC) in general, can be stated as in equation 2.1.

$$v(s) = (x(s), y(s)) \quad 0 \leq s \leq 1 \quad (2.1)$$

This behavior is achieved by either shrinking or expanding the curve, 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 2.2.

$$E_{snake} = \int (E_{internal}(v(s)) + E_{external}(v(s))) ds \quad (2.2)$$

The energy terms in equation 2.2 can be modified in order to represent the contour as a set of 2D control points and therefore convert into the discrete domain as given by equation 2.3 & equation 2.4.

$$v_i = (x_i, y_i) \quad 0 \leq i < n \quad (2.3)$$

$$E_{snake} = \sum_{i=0}^{n-1} (E_{internal}(v_i) + E_{external}(v_i)) \quad (2.4)$$

The internal energy ($E_{internal}$) stated in equation 2.4, consists of two terms which govern the motion of the snake when it is under no influence from the data term. As depicted in equation 2.5 (by discrete approximation), these two terms are respectively stated in literature as the elasticity and the stiffness term which ensures 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 according to the application in order to decide the contribution of each energy term towards the motion.

$$E_{internal} = \sum_{i=0}^{n-1} \alpha |v_{i+1} - v_i|^2 + \beta |v_{i+1} - 2v_i + v_{i-1}|^2 \quad (2.5)$$

External energy component in equation 2.4 was originally defined to incorporate the edge information of the image and to repel the motion enforced by the internal energy component near object boundaries. Equation 2.6 shows this interpretation where the term $\nabla I(v_i)$ defines the edge strength at control point v_i of the image I . Parametric snakes are in general solved as an energy minimization problem. Yet, as for the external energy (data term energy), it is desirable to maximize this value. Therefore a negative sign is used in equation 2.6 to correspond with the general energy minimization framework. Furthermore, γ is the scaling factor used to balance the external and internal energies to prevent the snake from missing edge points and ultimately shrinking to a single point.

$$E_{external} = - \sum_{i=0}^{n-1} \gamma |\nabla I(v_i)|^2 \quad (2.6)$$

Going back to the parametric representation (equation 2.2), it can be further shown that a snake which minimizes energy should satisfy the Euler-Lagrangian equation shown below.

$$\alpha v''(s) - \beta v''''(s) - \nabla E_{external} = 0 \quad (2.7)$$

This can further be expressed as a dynamic equation by treating $v(s)$ as a function of time (t) [34],[35],

$$v_t(s, t) = \alpha v''(s, t) - \beta v''''(s, t) - \nabla E_{external} \quad (2.8)$$

The above equation will be used to represent the GVF snake energy terms which will be discussed in the subsequent section 2.3.2.

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

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

Yet, the basic snake model described by Kass [27] suffers from the following drawbacks,

- Snakes are not guaranteed to find the global solution for the problem and often converges to a local minima depending on the intialization.
- The basic parametric active contour model cannot handle topological changes in the object of interest.
- The snake contours have the possibility to twist (fold on each other) which is highly unlikely to be present in real objects.
- This model also has very limited capture range for the data term and thus has problems in negotiating concave boundaries.
- Parametric active contours are highly susceptible to image noise that is especially present in the edge map.

Among the mentioned 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 address the limitation in 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 and this can be represented as in equation 2.9. Yet, this application also blurs the image boundaries and the exact positioning of the end segmentation result depends on the variance (σ) value of the Gaussian distribution used [36]. If the variance (σ) is set to a higher value, the capture range will be larger. But this will also produce highly blurred image boundaries which can adversely affect the accuracy of the final contour positioning.

$$E_{external} = - \sum_{i=0}^{n-1} \gamma |\nabla(G_{\sigma} * I(v_i))|^2 \quad (2.9)$$

Another common approach is to apply the standard distance transform (DT) to the image of interest. Here, a distance map is created using intensity edges as the feature points. The distance value is set to be proportional to the shortest distance from any of these feature points. Equation 2.10 illustrates the use of distance transform as an external energy in parametric snakes. Note the absence of the negative sign in equation 2.10 compared to the external energy functions given by equation 2.6 & 2.9. This is because the value of $D(v_i)$ reduces as the point becomes closer to image boundaries (feature points).

$$E_{external} = \sum_{i=0}^{n-1} D(v_i) \quad (2.10)$$

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

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

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

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

Cohen [37],[38] applied a non-linear transformation to the above mentioned distance maps to further enhance the capture range of a snake. These forces, which were referred to as 'distance potential forces', only altered the magnitude of the potential forces acting on the edge map while retaining the original vector field orientation.

2.3.2 Gradient vector flow as an external energy

As mentioned above, the use of the conventional snake model discussed has two main limitations when applying to real world segmentation problems. The first is the high sensitivity to the initialization of the snake control points with respect to the data terms present in the image. For an 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 and would miss the actual boundary. Also, at boundary concavities, the direction of the image gradient (on each side of the concavity) would point in opposite direction and avoids the snake from converging toward concave regions (see figure 2.3). Therefore, given the edge map ($\text{edge}(I)$) such that,

$$\text{edge}(I_{x,y}) = E_{\text{external}}(x,y)$$

we can define a static vector field $v(x,y) = [u(x,y), v(x,y)]$, which minimizes the energy functional [34] given below,

$$\varepsilon = \iint_{x,y} \mu \left(u_x^2 + u_y^2 + v_x^2 + v_y^2 \right) + |\nabla \text{edge}|^2 |v - \nabla \text{edge}|^2 dx dy \quad (2.11)$$

where ∇edge is the gradient of the edge map and u_x is the partial derivative of component $u(x,y)$ with respect to x ($u_x = \frac{\partial u(x,y)}{\partial x}$). A close observation of the equation 2.11 reveals the following two behaviors of the energy functional ε ,

- At coordinate locations where $\nabla \text{gradient}$ is small (homogenous regions), the functional ε is influenced by the partial derivatives of the vector field. This

ensures a smooth variation along homogenous regions where the traditional energy models would have no response. Therefore, this first term of equation 2.11 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 and the contribution of the second term of equation 2.11 increases. The minimum value (0) for the energy functional in equation 2.11 is achieved by setting $v = \nabla gradient$ around the vicinity or object boundaries and therefore preserves the conditions for a fast convergence.

Then by replacing the external energy component ($E_{external}$) in equation 2.8 with v (obtained by minimizing equation 2.11), the following representation can be achieved,

$$v_t(s, t) = \alpha v''(s, t) - \beta v''''(s, t) - v \quad (2.12)$$

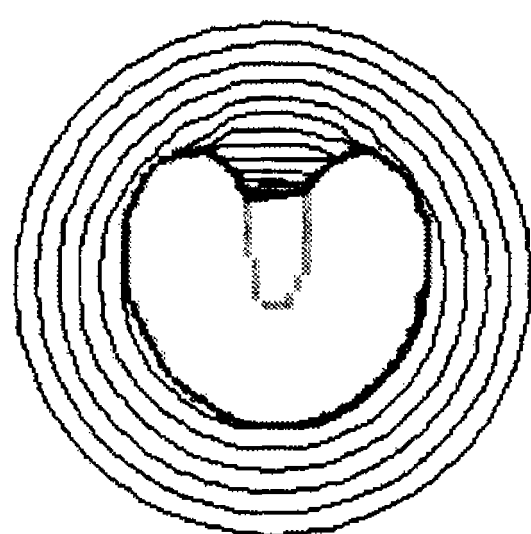
The parametric curve obtained by solving this equation 2.12, is called a 'GVF snake'. Further details of solving equation 2.12 can be found in [34],[35],[36].

2.3.3 GVF snakes vs DT snakes

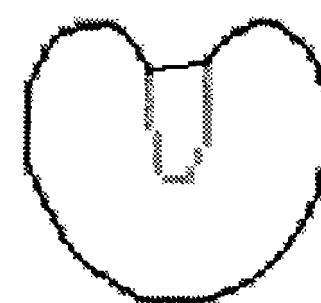
In this section, we will examine image results of applying two different external energy models to active contours, based on the image and code examples provided by Prince & Xu [2]. Figure 2.2 provides such a result depicting the difference in convergence of the DT based snake with the GVF snake. Figure 2.2(a) and 2.2(c) elaborate that the GVF converges faster and more deeper into the concave region of the image and the final result of the DT snake (figure 2.2(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 2.3 in which the following observations can be made,

- The GVF model vector field is more dense relative to the DT snake field. The GVF field is specifically stronger near object edges and decreases (in magnitude) slower than the DT model, when going away from these boundaries. This observation explains the faster convergence.

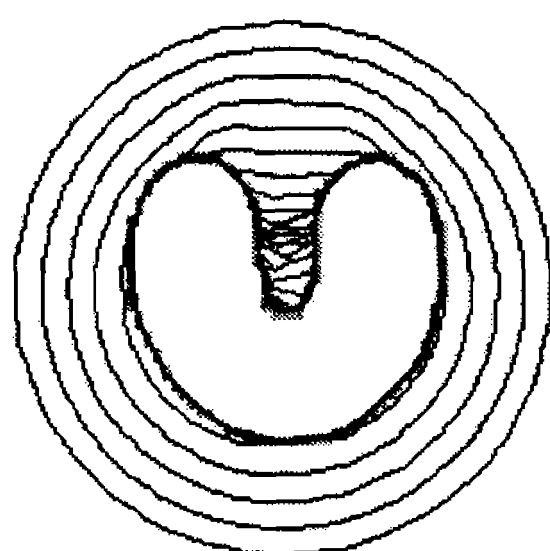
- The GVF field has the property of pointing towards the concave boundary (in the mid section of the U shape concavity), while the DT model vectors simply exert forces with opposite directions (in the same region). This cancels out the influence of the external energy term in the DT model and therefore, the snake stops traversing towards the concave region.



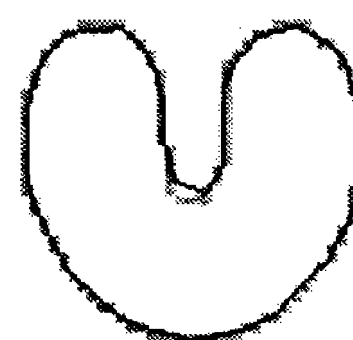
(a) DT model iterations



(b) DT model final result



(c) GVF model iterations



(d) GVF model final result

Figure 2.2: 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) [2].

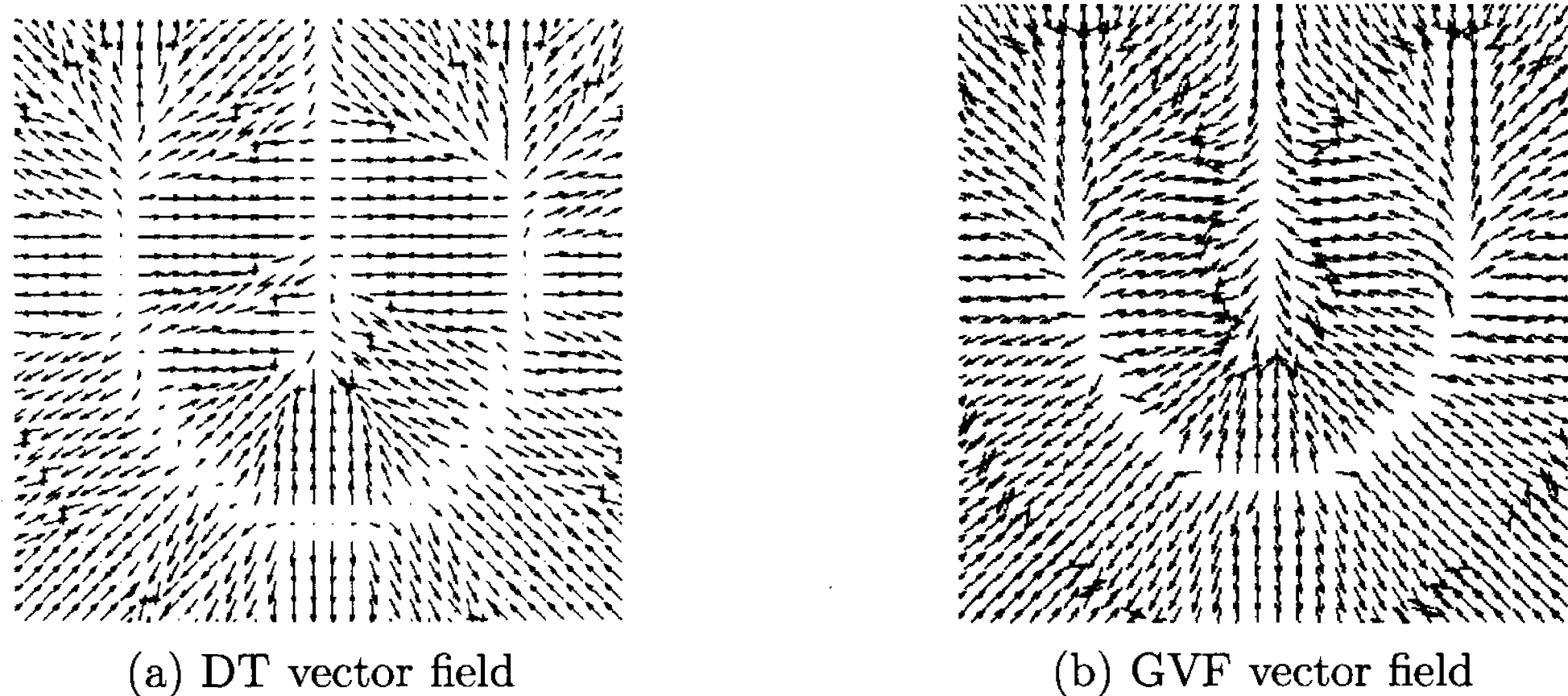


Figure 2.3: Comparison between Distance Potential (DT Based) vector field (left) and the GVF vector field (right) [2].

2.4 Discrete curve evolution (DCE)

DCE is a methodology that can be adopted for evolving polygons while preserving visual information which can yield a hierarchical set of polygons according to their significance in representation of the original object. In other words, DCE is an effective and robust tool for generalizing polygonal contours based on digital linearization [3]. DCE 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:

- DCE can be directly used for shape simplification which can then be used for comparing different shapes. A detailed study of using DCE can be found in one of the publications of Latecki & Lakämper [39] while some image examples can be found on web resources [40]
- DCE can also be used for object extraction from a database when a query is given in the form of a visual sketch. In this scenario, 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 [41].

- The polygons which are obtained through the DCE process can be effectively used for skeleton pruning as described by Bai et al [42]. The DCE result is 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. This aspect of DCE is used in our research for obtaining the centerline of a chromosome.

2.4.1 Definitions

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

- $v(P^i)$ as a vertex contained in the polygon P^i
- $arc(s_i, s_{i+1})$ 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 . The equation used for calculating this relevance measurement and the rationale behind using it is explained in section 2.4.2. The algorithm of discrete curve evolution by digital linearization is illustrated below [3],[43],[44],

The DCE algorithm:

1. find the value of,

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

2. and then 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,

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

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

$$V(P^{i+1}) = V(P^i) \setminus V_m(P^i)$$

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

This new polygon creation involves 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 performed for any desired termination criterion, for an example, until the end polygon becomes convex.

The end criterion has to be set up according to the application and the end result convexity can be assured by stopping the process at a higher stage of evolution. The convexity of the obtained polygonal partitions is important as convex shapes determines the visual parts of an object [3]. If the stopping criterion is inappropriate, the algorithm will converge to a degenerate solution of a polygon $P = \{\emptyset\}$. The shape simplification process and the immunity to noise of the above defined function can be clearly seen in figure 2.4². Some feature points are marked in figure 2.4 to show the stability with noise deformations and to show the similarity of the two evolution results.

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

2. and then 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,

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

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

$$V(P^{i+1}) = V(P^i) \setminus V_m(P^i)$$

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

This new polygon creation involves 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 performed for any desired termination criterion, for an example, until the end polygon becomes convex.

The end criterion has to be set up according to the application and the end result convexity can be assured by stopping the process at a higher stage of evolution. The convexity of the obtained polygonal partitions is important as convex shapes determines the visual parts of an object [3]. If the stopping criterion is inappropriate, the algorithm will converge to a degenerate solution of a polygon $P = \{\emptyset\}$. The shape simplification process and the immunity to noise of the above defined function can be clearly seen in figure 2.4². Some feature points are marked in figure 2.4 to show the stability with noise deformations and to show the similarity of the two evolution results.

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

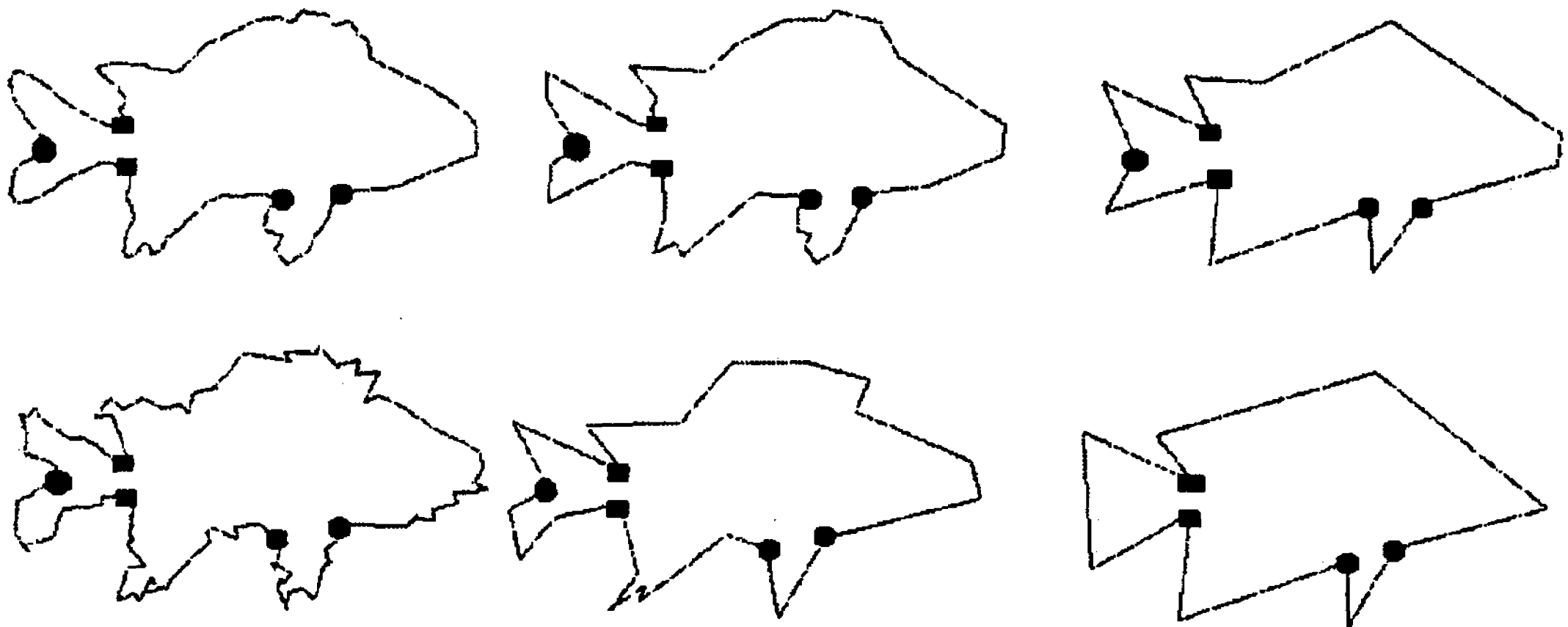


Figure 2.4: above: Steps of the DCE process of shape simplification, below: the same steps when initiated by adding boundary noise to the same image [3].
(Reproduced with permission from Prof. Longin Jan Latecki)

2.4.2 The relevance function

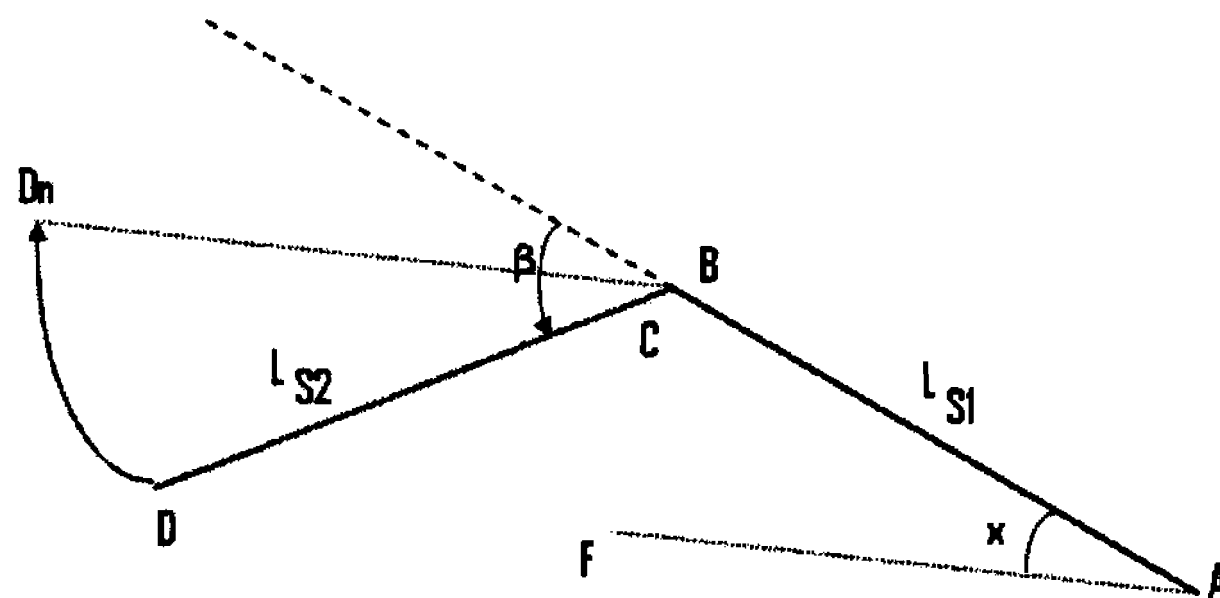


Figure 2.5: A representation of two line segments and the used angle measurements in the DCE process, which can be used to explain the rationale behind the used 'relevance measure'.

The idea of discrete curve evolution is to obtain a hierarchical set of polygons that represent the shape features of the original contour $C \in R^2$. The effectiveness of this contour evolving process depends on the measurement that is 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 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 rationale behind the above assumption is highly intuitive and can be explained using figure 2.6. But first, a brief description of the local features and their notations that can be used for this explanation is in order. Lets take line segments A-B & C-D as 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 is given by $\beta(s_1, s_2)$ as illustrated in figure 2.5, which is given by $|angle(C - D) - angle(A - B)|$. Then referring to figure 2.6, we can explore the effects of the assumption made earlier. The contour segments $C2$ and $C4$ are equal in length and shape, and the only difference is in the shape and length of $C3$ and $C4$ arcs. The shape contribution of the arc $C1$ is obviously higher than that of $C3$ with respect to rest of the contour. Its also evident that the turn angle and the length of the segments of $C1$ is higher than $C3$. Therefore, the assumption made earlier regarding the relevance measure dependencies can be justified from these observations. In order to get the global perspective, the lengths of the segments are normalized with respect to the total length of the contour/polygon.

Next, a suitable function has to be formulated to reflect the previously mentioned two parameters. This is achieved by 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) can be found by getting the difference of 'y' axis values between two consecutive entries. Then, an angle $\angle x$ ($0 \leq x \leq \beta$) 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 2.5). This angle $\angle x$ can be represented as $\angle x = [\beta(s_1, s_2) \times L_{s2}] / [L_{s1} + L_{s2}]$ where both L_{s1} and L_{s2} are normalized lengths of the segments [3]. Then, the circular arc-length ($L_{s1} \times x$) is defined as the relevance function for the DCE process as follows,

$$K(s_1, s_2) = \frac{\beta(s_1, s_2) \times L_{s1} \times L_{s2}}{(L_{s1} + L_{s2})} \quad (2.13)$$

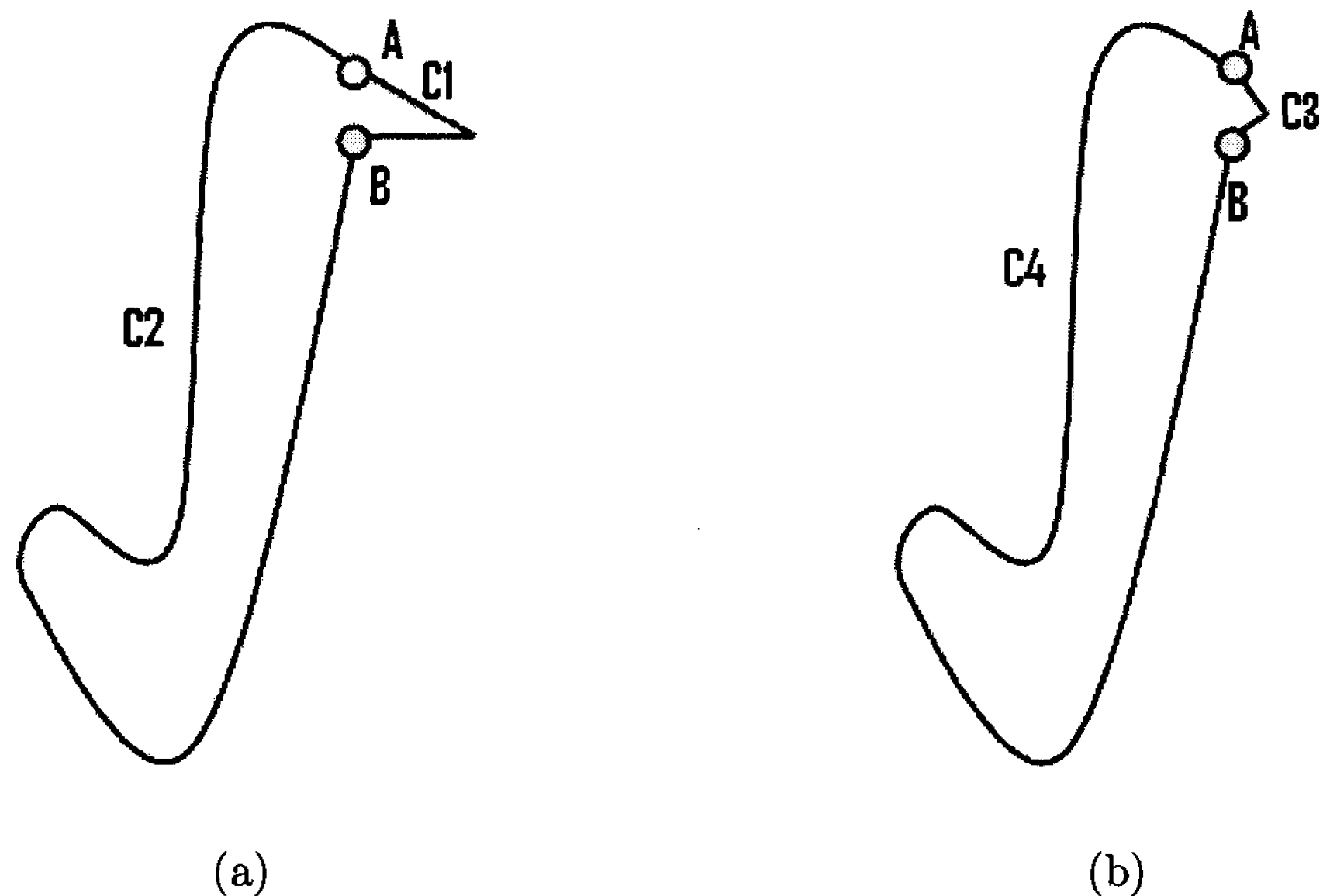


Figure 2.6: Shape variations of polygon partitions and the effects of turn angle and arc length to the relevance measure.

The above relevance value can be considered as the linearization cost for a given arc of the polygon. Also, due to normalization of length values, this relevance measure has a global representation although its calculated locally. Further explanation regarding the derivation of the equation 2.13 and the tangential space representation can be found in [3] along with some image examples at [40].

2.4.3 Advantages & disadvantages of DCE

The discrete curve evolution based on digital linearization has the following advantages compared with other existing methods for shape simplification [43],[3],[44] (one of the main comparison method is the shape simplification work carried out by Siddiq and Shokoufandeh [45]).

- DCE method is rotation, reflection, translation and scaling invariant. The rotational invariance is due to the use of the tangent space for the polygon evolution process.
- Unlike other methods which are based on local extremal points, DCE is robust in real world discrete digital images. It is also robust against boundary noise

(digitization errors) in digital images and removes them 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 also implies stability against noise and is proved in [44].

- This algorithm is guaranteed to converge as at least one vertex is deleted in every iteration.
- The evolution 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 respect to the whole contour.
- The resulting polygons at different stages of the algorithm yield different relevance levels which can be taken together to form a relevance hierarchy of polygonal representations. Therefore, the higher the stage of contour evolution, the higher is its relevance to the original object.
- The initial polygon used for the evolution does not need to be simple. Therefore, 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 .
- DCE method does not introduce any blurring to the object boundaries and also does not dislocate any relevant features.

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.

2.5 Cubic spline interpolation

Given a set of initial points (x & y coordinates), it is often necessary in numerical analysis, to come up with a suitable method to generate the points in between and make a smooth distribution. In the field of numerical methods, the algorithms for deriving these points are called 'interpolation algorithms'.

Suppose the following x and y ($= f(x)$) values are given,

$$\begin{aligned} x &: x_0, x_1, x_2, \dots, x_{n-1}, x_n \\ y &: y_0, y_1, y_2, \dots, y_{n-1}, y_n \end{aligned}$$

Then, 'interpolation' can be defined as the technique of estimating the values of a function for any intermediate value of the independent variable (i.e getting all data points y for any x within $x_0 \leq x \leq x_n$) [46]. The term 'splines' was initially used for springy pieces of wood/steel which were used to make curves in railway design and construction. Splines in mathematics, represents a piece-wise combination of polynomial functions that can be used to smoothly represent a given set of data points. By combining these polynomials while preserving continuity allows splines to closely represent a given set of data points while avoiding the Runge's phenomenon [47] when using high-degree polynomials. The above mentioned Runge's phenomenon is the main limiting factor for the use of polynomial approximation. The term 'cubic' implies that the polynomials used are of 3^{rd} degree. Therefore, cubic spline interpolation represents a given set of data points by piecewise 3^{rd} degree polynomials and this was first introduced by Ferguson [48],[49].

A curve segment can be represented in one of the following three methods [50],

- **explicit representation** : The mathematical function can be stated as $y = f(x)$. Although this is one of the most simplest forms of mathematical representation, the scope of functions covered by this is limited. For an example, this cannot represent vertical lines ($x = Const$) and thus demands the x values of the curve to be distinct [51].

- **implicit representation** : This represents the function in the form of $F(x, y) = 0$ and this notation can define any known function between x & y unlike the explicit representation.
- **parametric representation** : The parametric representation is used mainly when presented with a collection of data points with an unknown underlying function. Here, the parametric curve (PC) can be represented as $P(t) = (x(t), y(t))$, where t usually lies in the range of $[0, 1]$.

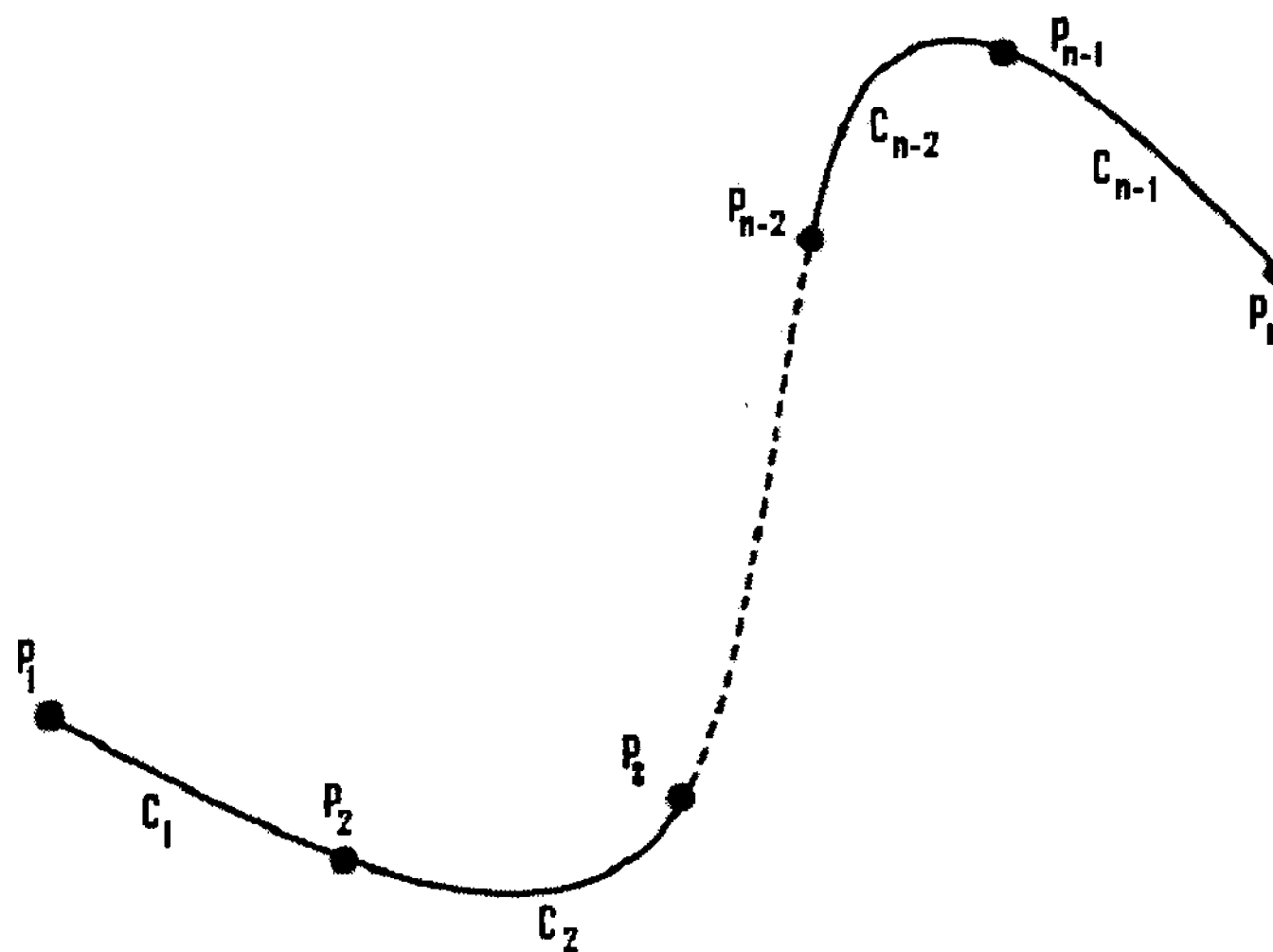


Figure 2.7: Spline curve segment arrangement

2.5.1 Definitions

When given n number of data points (obtained from the parametric curve - $P(t) = (x(t), y(t))$ over x axis), labeled from P_1 to P_n , the cubic spline $S(x)$ can be represented from the following collection of $(n - 1)$ functions (refer figure 2.7),

$$S(x) = \begin{cases} C_1(x), & P_1 \leq x \leq P_2 \\ C_2(x), & P_2 \leq x \leq P_3 \\ \vdots \\ C_i(x), & P_i \leq x \leq P_{i+1} \\ \vdots \\ C_{n-1}(x), & P_{n-1} \leq x \leq P_n \end{cases}$$

where each C_i is a cubic polynomial function with the form:

$$C_i(x) = a_i x^3 + b_i x^2 + c_i x + d_i \quad (2.14)$$

a_i, b_i, c_i and d_i are real value constants ($\in \mathbb{R}$) that need to be determined. Therefore, this yields $4 \times (n - 1)$ unknown parameters. In order to find a unique solution for the spline, a $4 \times (n - 1)$ number of independent equations relating the above parameters are needed. In interpolation (unlike in approximation), the curve has to pass through all given data points ($P_1 - P_n$) and this yields $2 \times (n - 1)$ equations. In order to use the $y = f(x)$ (explicit) representation for each curve segment, the following provisos have to be met [51],

- All the points within each segment should have distinct x values.
- All the points are represented in the ascending order of x coordinates.

Cubic splines require the first derivatives at the $(n - 2)$ interior points (points connecting each curve segment) to be the same. This is interpreted in equation 2.15.

$$C_i'(P_{i+1}) = C_{i+1}'(P_{i+1}), \quad 1 \leq i \leq (n - 2) \quad (2.15)$$

where $C_i'(P_i)$ is the first derivative of the curve segment C_i at point P_i . Considering the polynomial equation 2.14, this can be written as,

$$3a_i P_{i+1}^2 + 2b_i P_{i+1} + c_i = 3a_{i+1} P_{i+1}^2 + 2b_{i+1} P_{i+1} + c_{i+1}$$

For better continuity, Cubic spline interpolation further demands the second order derivatives to be same at the interior points. Equation 2.16 demonstrates this

condition where $C_i''(P_i)$ would yield the second derivative of C_i curve segment at point P_i .

$$C_i''(P_{i+1}) = C_{i+1}''(P_{i+1}), \quad 1 \leq i \leq (n - 2) \quad (2.16)$$

This can be similarly stated as:

$$6a_i P_{i+1} + 2b_i = 6a_{i+1} P_{i+1} + 2b_{i+1}$$

The above two conditions for the continuity of the complete spline curve provides $2 \times (n - 2)$ equations and summing up to a total of $(4n - 6)$ equations, which leaves a deficit of two equations (for a unique solution). These two equations are generated from the 'boundary (end) conditions' specified for the spline curve, which is discussed in section 2.5.2.

2.5.2 Spline boundary conditions

There are numerous end (boundary) conditions in literature for filling up the above mentioned deficit of equations for solving the spline interpolating curve [52],[53].

Clamped end condition: Here, the user has to provide values for the first derivatives of the two end points of the curve ($C_1'(P_1)$ & $C_{n-1}'(P_n)$). This method of end point selection is highly user interactive and the shape of the curve at both end points completely rely on the initial presumption of the user. If the user inputs are K_1 and K_2 for $C_1'(P_1)$ and $C_{n-1}'(P_n)$ respectively, then the two new equations are,

$$\begin{aligned} 3a_1 P_1^2 + 2b_1 P_1 + c_1 &= K_1 \\ 3a_{n-1} P_n^2 + 2b_{n-1} P_n + c_{n-1} &= K_2 \end{aligned}$$

and the system can be easily solved using these $4 \times (n - 1)$ equations.

Relaxed end condition: In this method, it is assumed that the two extreme end points have a zero curvature value. This boundary condition restricts the spline curve from having drastic changes of directions at its extreme end points. This

method is also referred as the 'natural condition' in literature [50]. The resulting equations are as follows,

$$\begin{aligned}6a_1P_1 + 2b_1 &= 0 \\6a_{n-1}P_n + 2b_{n-1} &= 0\end{aligned}$$

Cyclic condition: This boundary condition is more suitable for splines that forms cycles by joining together at the ends and therefore, the condition is that the first derivatives and the second derivatives of the two extreme points must be same. This can be viewed as an extension of equation 2.15 and equation 2.16 to include its extreme points. That is,

$$\begin{aligned}3a_1P_1^2 + 2b_1P_1 + c_1 &= 3a_{n-1}P_n^2 + 2b_{n-1}P_n + c_{n-1} \\6a_1P_1 + 2b_1 &= 6a_{n-1}P_n + 2b_{n-1}\end{aligned}$$

Anti-cyclic condition: This boundary condition states that the first and the second derivative of the first point has to be equal to the 'negative' of the first and the second derivatives of the last point (other extreme point) of the curve. Therefore, this method also behaves closely with the cyclic condition.

$$\begin{aligned}3a_1P_1^2 + 2b_1P_1 + c_1 &= -1 \left(3a_{n-1}P_n^2 + 2b_{n-1}P_n + c_{n-1} \right) \\6a_1P_1 + 2b_1 &= -1 (6a_{n-1}P_n + 2b_{n-1})\end{aligned}$$

Not-a-knot condition: This condition was first introduced by De Boor [54] and is commonly used even at present. Here, the third derivatives at the first interior point (P_2) and the last interior point (P_{n-1}) are considered same for both curve segments. Therefore, this boundary condition provides a smoother curve as the output. This is the boundary condition used in this dissertation for spline interpolation in section 3.2. The two simple equations obtained here are,

$$\begin{aligned}a_1 &= a_2 \\a_{n-2} &= a_{n-1}\end{aligned}$$

Chapter 3 Proposed algorithm

The use of FISH probes in fluorescence microscopy has been rapidly increasing as it marks a specific location (with known coordinates) in the human genome. This information can be directly adopted for more accurate diagnosis of many genetic diseases. In order to speed up this clinical diagnosis process, computer based applications are highly essential. Yet, most of the available techniques have been developed for karyotype analysis, which is performed on G - banded chromosomal images. On top of that, most of these existing methods have the following limitations,

- They work effectively only on straight or slightly bent chromosomes. Yet, in reality chromosome shapes are highly variable in shape (on microscope slides) and therefore, the above limitation is not satisfactory. These sharp bends are also evident in the pro-metaphase stage of cell division (mitosis).
- The extracted centerline of the chromosomes in these methods can bear spurious branches and a satisfactory pruning approach is not available. These unwanted branches could be a direct influence of boundary noise present in digital images and therefore need to be addressed.

Therefore, this chapter will discuss our proposed algorithm developed to overcome the above mentioned limitations in literature. This algorithm was developed and tested on DAPI stained metaphase and pro-metaphase cell images, but can well be adopted for Geimsa or any other banded chromosomal images. In our approach, we have paid more interest in dealing with long and highly bent chromosomes which were excluded in other existing methods due to the high complexity and variability of shape. This proposed method is based on two basic assumptions:

1. It is assumed that the chromosome of interest is not overlapping or touching other chromosomes in the same cell image. This is a reasonable assumption due to the screening provided by the ranking algorithm [16] discussed in section 2.1.

2. It is assumed that the coordinate values of the probe signal is given before hand. A set of filters are currently being developed for this purpose and will be further discussed under future work. (see section 5). The proposed algorithm was developed and tested by manually representing those probe signals as high intensity pixels at the probe locations.

The proposed algorithm is explained in six functional stages, titled as following,

- Pre-processing and segmentation
- Centerline detection
- Chromosome end point detection
- Centromere identification and polarity assignment
- FISH probe projection
- Fractional ratio measurement

3.1 Pre-processing and segmentation

First of all, as in many image processing applications, a proper pre-processing stage was required in order to achieve the expected results from the algorithms at later stages. The pre-processing algorithm that we have followed is depicted in figure 3.1 as a flow chart. In this method (unlike in traditional karyotyping), at any given time we were solely interested in a particular homologue chromosome pair rather than the whole 23 pairs which were most likely present in the chromosome cell image. Therefore, a proper method for detecting the probe signals from the DAPI (4',6-diamidino-2-phenylindole) stained images was required for the full automation of the abnormality detection algorithm. Figure 3.2 depicts such a DAPI stained digital image with highly distinguishable multiple probe signals (red dots) on chromosome 15 (labeled in the image).

Fluorescence chromosome images tend to have the following characteristics,

- The stored images in the computer, are RGB color images. From an image processing point of view, these RGB colors carry very little importance in analyzing these chromosomes.

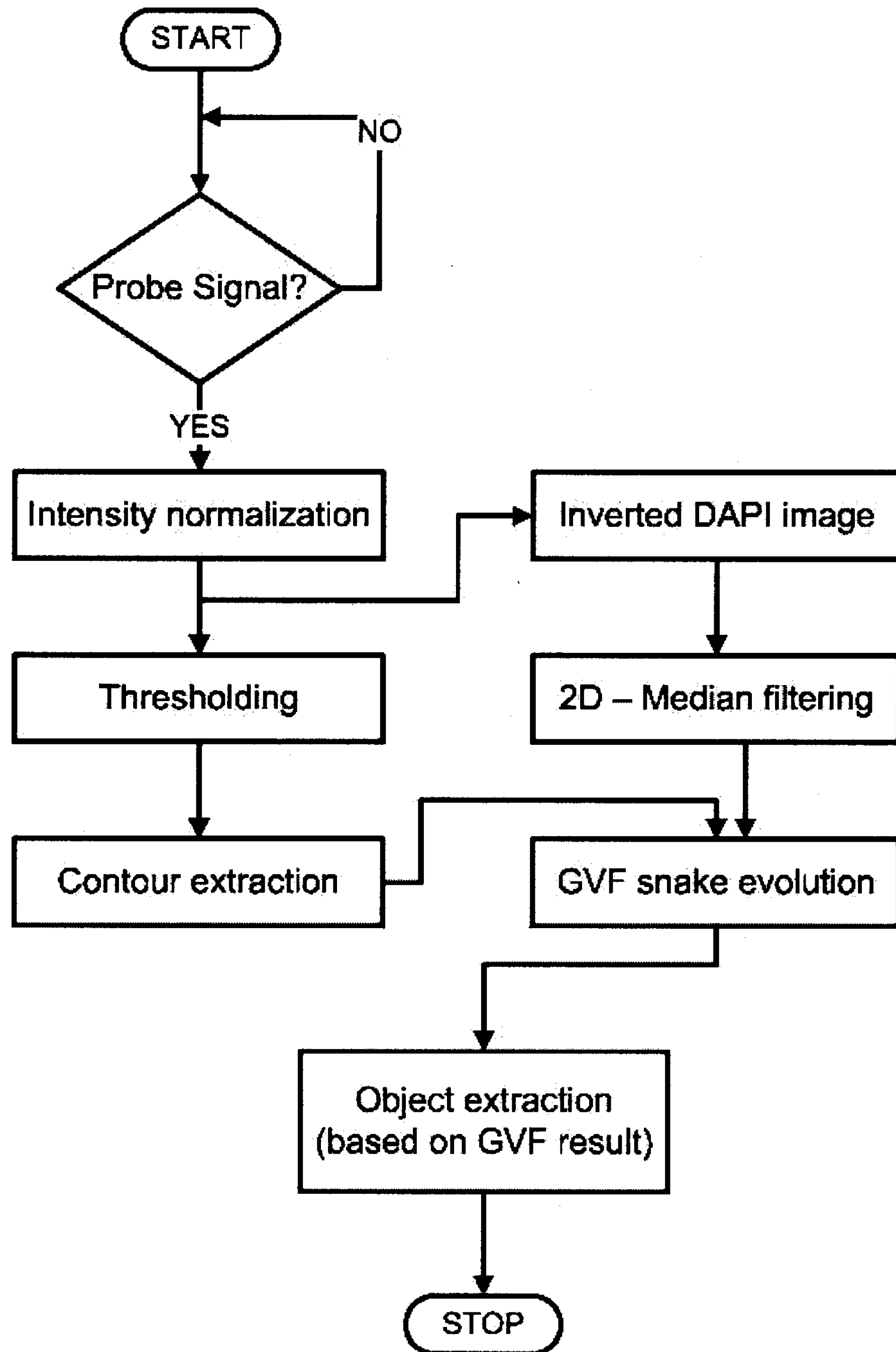


Figure 3.1: The flow chart of the Pre-processing stage algorithm.

- Especially the background pixels can have highly variable intensity values depending on its placement in the fluorescence image mainly due to light source effects.
- These images tend to have a limited range of intensities (out of the possible 255 levels in this case). The effects introduced by the fluorescence light source also contributes to this feature. For example, the middle of a fluorescence microscopy

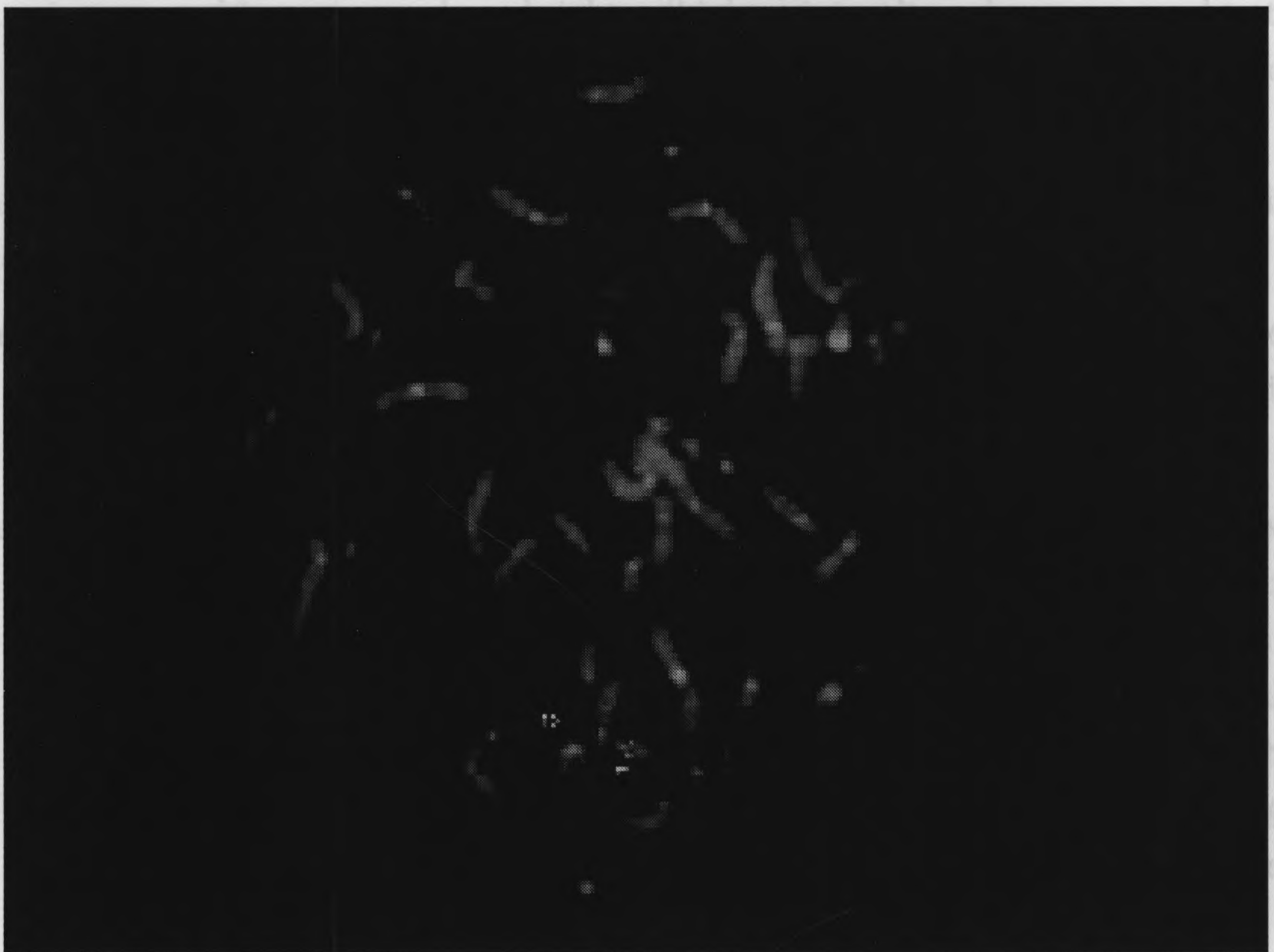


Figure 3.2: A digital cell image with two FISH probes located on chromosome 15 and its homologue. This also depicts the difficulty (image processing wise) in locating these small probe signals in a chromosome cell image.

Then an intensity mapping was performed according to figure 3.3 where the window size was equal to the intensity range of the input image and the center been the center of the window. Equation 3.1 gives the point processing equation used for the contrast enhancement (intensity normalization), where f_{max} & f_{min} are respectively the highest intensity and the lowest intensity in the input image histogram.

$$I_{norm} = \frac{(I_{int} - f_{min})}{(f_{max} - f_{min})} * (255 - 0) \quad (3.1)$$

- Especially the background pixels can have highly variable intensity values depending on its placement in the fluorescence image mainly due to light source effects.
- These images tend to have a limited range of intensities (out of the possible 255 levels in this case). The effects introduced by the fluorescence light source also contributes to this feature. For example, the middle of a fluorescence microscopy image would on average have brighter (higher) intensity values compared to the corners of the image. Also, in terms of segmentation, a well spread bi-modal histogram intuitively would lead to better results.

Therefore, the following steps were formulated to address the above characteristics, prior to segmentation.

Step 1 First, the RGB fluorescent image was converted in to gray scale space (intensity range 0 - 255). This drastically simplifies the sub-sequent image processing steps while preserving information.

Step 2 In order to address the pixel intensity variation issue discussed before, a rectangular shaped window from the original DAPI image (with fixed dimensions and the FISH probe location as the center) was extracted. This window has to completely include the chromosome of interest (with the FISH probe) while also including some portion of the background as well. The dimensions of this has to be set manually prior to the analysis. Therefore, by effectively capturing a window, the effects of variable lighting is minimized and the processing becomes local.

Step 3 Next, the intensities of this extracted image portion was normalized. The normalization was performed through a process called 'window center adjustment' (see figure 3.3). First, the intensity range of the input image was detected. Then an intensity mapping was performed according to figure 3.3 where the window size was equal to the intensity range of the input image and the center been the center of the window. Equation 3.1 gives the point processing equation used for the contrast enhancement (intensity normalization), where f_{max} & f_{min} are respectively the highest intensity and the lowest intensity in the input image histogram.

$$I_{norm} = \frac{(I_{int} - f_{min})}{(f_{max} - f_{min})} * (255 - 0) \quad (3.1)$$

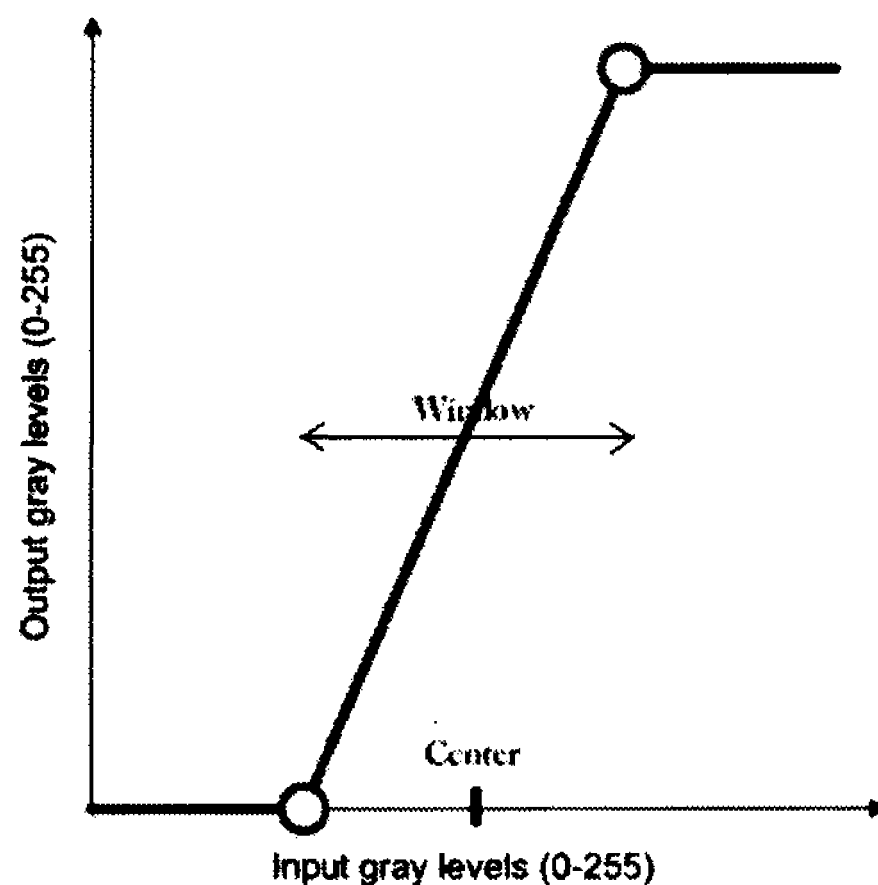


Figure 3.3: The window-center intensity mapping scheme which was used to map a certain intensity range (defined by the window and the center) to the full range (intensity levels 0 - 255).

Figure 3.4 illustrates the effects of the contrast enhancement explained above and how it made the intensities of the original dark image spread across the complete spectrum of possible values. Histogram equalization which is a commonly used histogram related pre-processing method, fails to yield satisfactory results for these images (see figure 3.4).

Next, a simple threshold was applied to the extracted image window based on Otsu's Method. Otsu's method is a clustering algorithm that attempts to find the optimum threshold value that minimizes intra-class variance (background/foreground). This intra-class variance σ_c is calculated according to equation 3.2, where $q_{c1}(t)$ & $q_{c2}(t)$ are estimated class probabilities while $\sigma_{c1}^2(t)$ & $\sigma_{c2}^2(t)$ 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 3.3.

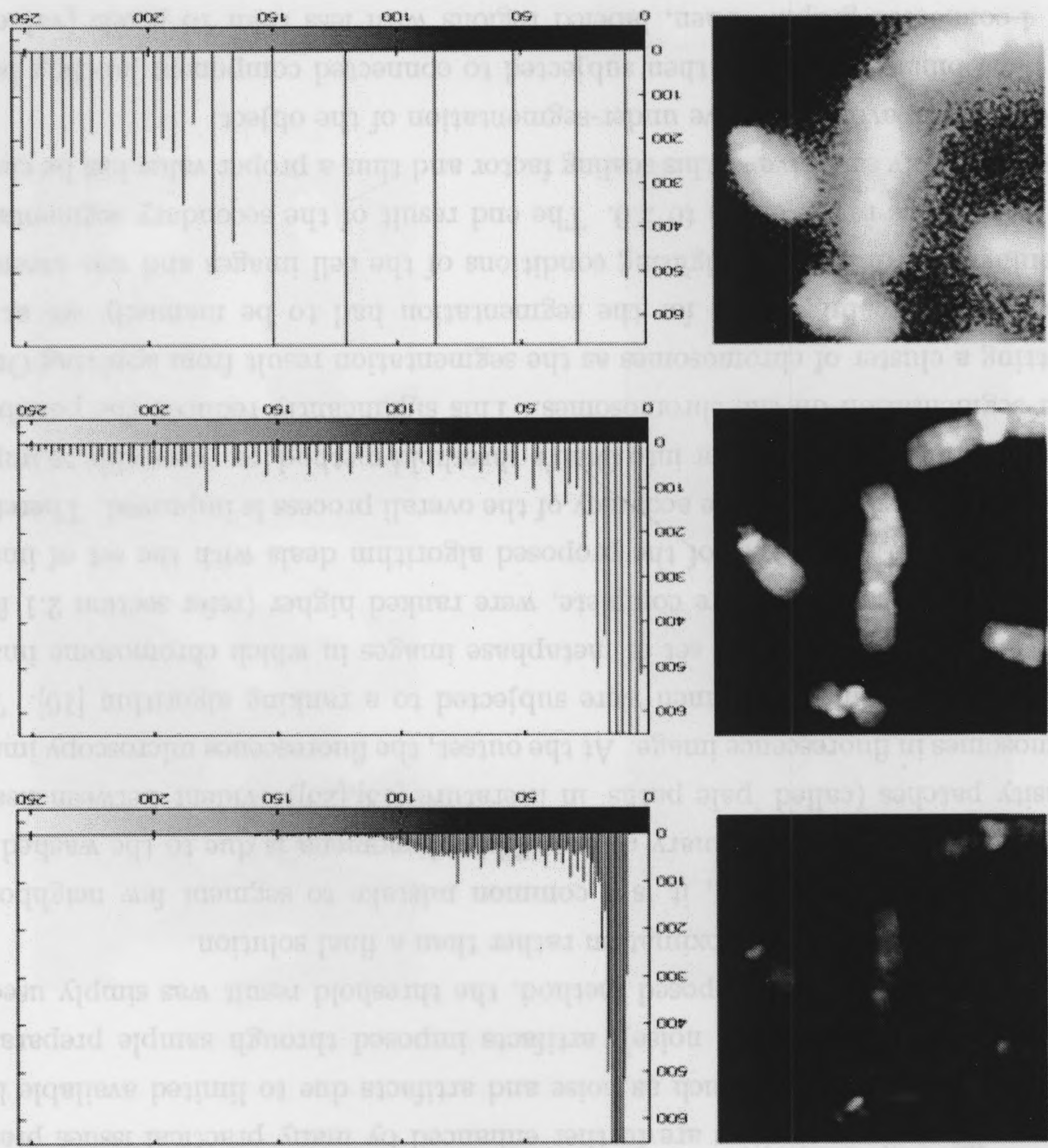
$$\sigma_c^2(t) = q_{c1}(t) * \sigma_{c1}^2(t) + q_{c2}(t) * \sigma_{c2}^2(t) \quad (3.2)$$

$$I_{bin} = \begin{cases} I_{bin} = 1, & \text{if } I_{x,y} \geq T_{int} \\ I_{bin} = 0, & \text{if } I_{x,y} < T_{int} \end{cases} \quad (3.3)$$

Thresholding has been adopted in many earlier works due to the unsophisticated nature of chromosome image histograms, which are mostly separable into two

non overlapping modes or classes. As explained by equation 3.3, this is a point processing method. That is, every point (pixel) decides individually without considering any spatial information. Therefore, this segmentation result inherits many imperfections. The threshold result was simply used as a common criteria to segment the neighboring pixels. The threshold result was simply used as a common criteria to segment the neighboring pixels. The threshold result was simply used as a common criteria to segment the neighboring pixels.

Figure 3.4: A comparison of the intensity histograms of the chromosome images with from top to bottom: Original cropped image, contrast enhanced image and the histogram equalized image.



non overlapping modes or classes. As explained by equation 3.3, this is a point processing method. That is, every point (pixel) decides individually without considering any spatial information. Therefore, this segmentation result inherits many imperfections. These irregularities are further enhanced by many practical issues present in fluorescence microscopy such as noise and artifacts due to limited available light (readout noise, photon shot noise), artifacts imposed through sample preparation etc.[55] Therefore in our proposed method, the threshold result was simply used as an initial segmentation approximation rather than a final solution.

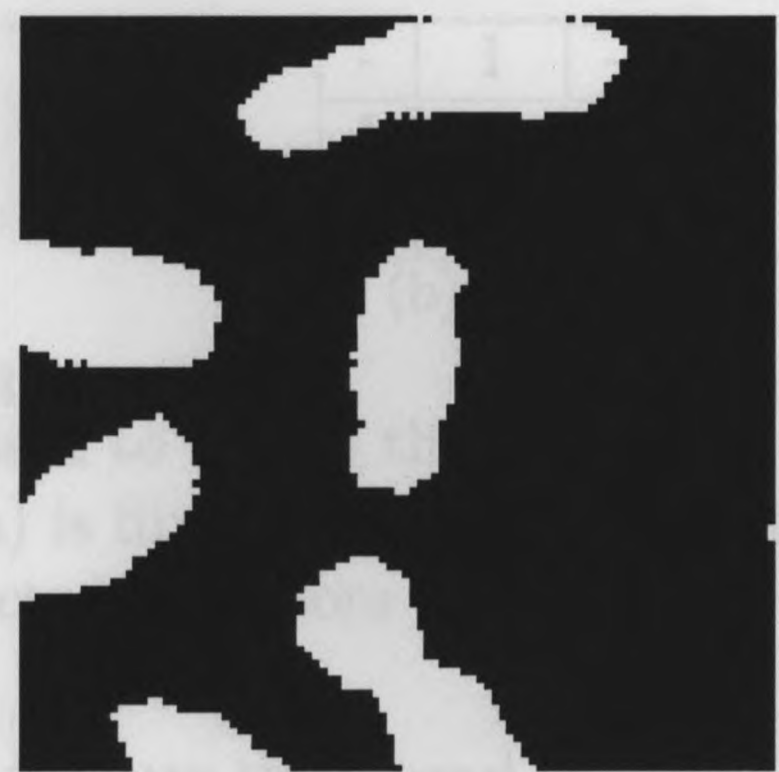
In binary thresholding, it is a common mistake to segment few neighboring chromosomes into a single binary object. This phenomena is due to the washed out intensity patches (called 'pale paths' in literature [13],[23]) evident between nearby chromosomes in fluorescence image. At the outset, the fluorescence microscopy images obtained from a single specimen were subjected to a ranking algorithm [16]. This algorithm provided a ranked set of metaphase images in which chromosome images that are 'well spread' and are complete, were ranked higher (refer section 2.1 for a brief description). The rest of the proposed algorithm deals with the set of images with the best rank so that the accuracy of the overall process is improved. Therefore, by introducing a scaling factor into Otsu's threshold method, we were able to impose under-segmentation on the chromosomes. This significantly reduced the possibility of getting a cluster of chromosomes as the segmentation result from applying Otsu's method. The scaling factor for the segmentation had to be manually set at the beginning according to the lighting conditions of the cell images and was meant to work within the range of 0.8 to 2.0. The end result of the secondary segmentation was significantly sensitive to this scaling factor and thus a proper value has be chosen in order not to avoid excessive under-segmentation of the object.

The binary image was then subjected to connected component labeling based on a 4-connected graph. Then, labeled regions with less than 10 pixels (value set empirically) were removed from the binary image. This assists in removing a lot of isolated noisy blobs in the thresholds result. Next, the object of interest (chromosome with the FISH probe) needed to be extracted from the segmented and labeled image. This was achieved by first getting the pixel coordinates of the FISH probe and by cross-checking with all labeled component coordinates. If a match is found, then all pixels with the same label (of the matching pixel) were extracted as the selected object. Yet, owing to the hard segmentation imposed earlier, the FISH probe coor-

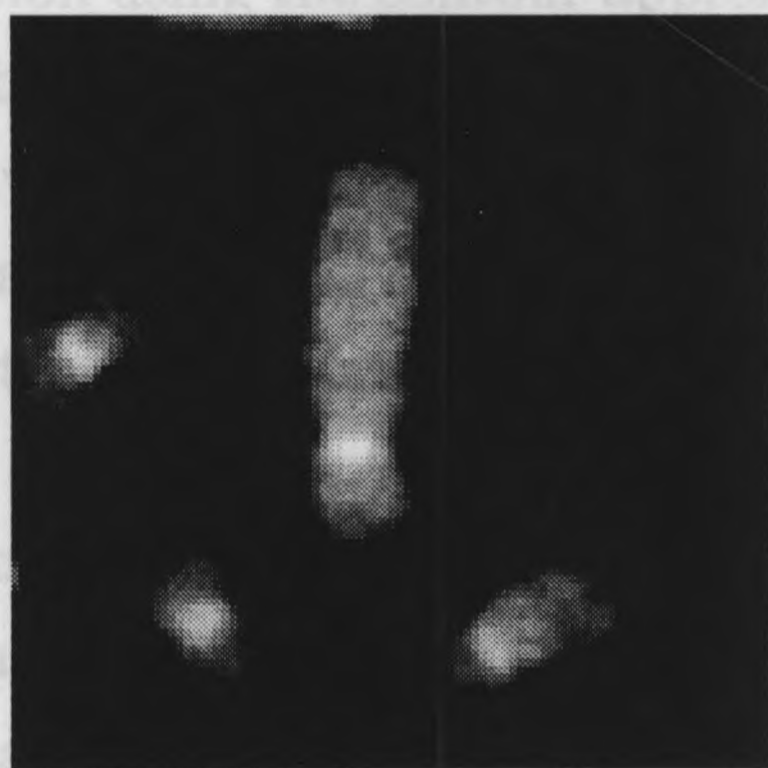
dinate may not match with any of the connected component labeled pixels. This is especially the case when the FISH probe is closer to the boundary of the chromosome. In such a case (no matches), the program was configured to pick the nearest labeled component to the FISH probe location. Figure 3.5 demonstrates some examples of the Otsu method application and the segmented results in this figure also justify the requirement for the nearest labeled component extraction process.



(a) Original image



(b) Thresholded result



(c) Original image



(d) Thresholded result

Figure 3.5: Examples of segmentation of two extracted image windows using Otsu method where the boundary discontinuities of the Otsu result are evident.

Next, the extracted binary object was subjected to a morphological filling operation where every pixel in the image with a value '0' which has 4-connected pixels with value '1' was complemented to have value '1'. This process was needed to remove any possible discontinuity that could be present both at the boundary as well as within the object. Since thresholding is a point processing scheme (given by equation 3.3), any noisy pixel in the object region can be grouped as background.

Having extracted the chromosome from the binary image, the contour of this object needed to be produced. The contour was needed merely to be used as the initial contour for the second stage of segmentation. Thus, the binary result was first subjected to a morphological process, in which a pixel's value was iteratively changed to zero (0) if all of its 4 connected neighbors are of value '1' as demonstrated in table 3.1. Next, the contour was traced using a (3×3) neighborhood [56].

-	1	-
1	1	1
-	1	-

(a) before

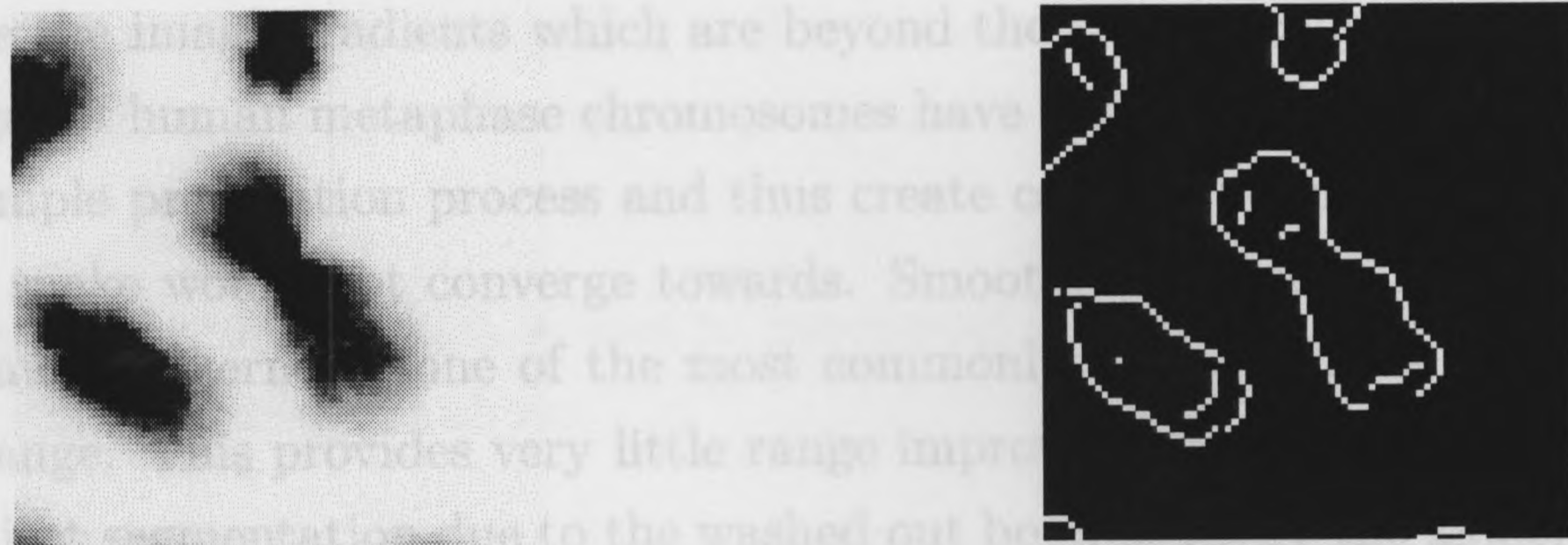
-	1	-
1	0	1
-	1	-

(b) after

Table 3.1: The 4 connected removal process used to obtain the contour of the selected object. Here, the middle pixel of 3.1(a) is flipped from '1' to a '0' as in 3.1(b) based on the 4-connected neighbors

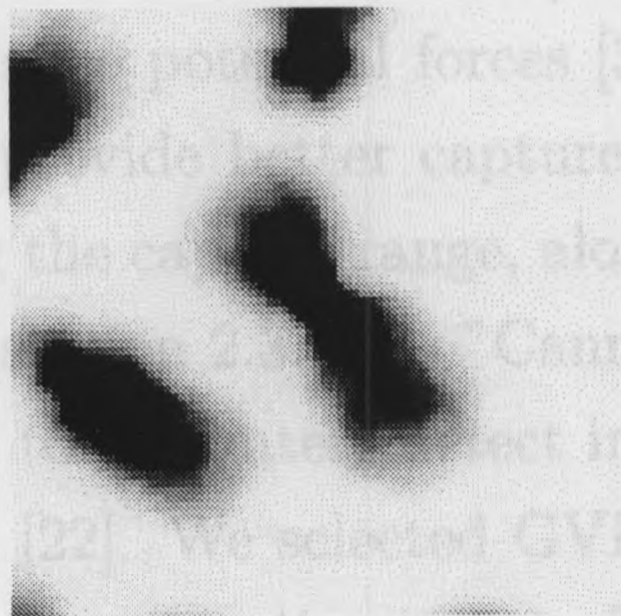
As depicted in figure 3.1, the next step is to perform the second stage of the segmentation using the contour approximation obtained from the previous stage. Therefore, the negative of the DAPI (inverted DAPI) image was taken for this purpose. The inverted DAPI is commonly used to illustrate a wide range of image intensity information like banding patterns. The image at this stage consisted of a significant amount of noise and thus had to be properly reduced through filtering in order to obtain an accurate segmentation. Figure 3.6 illustrates the effects of applying median filter, mean filter & Gaussian filter. Gaussian filter had a variance (σ) value of 3.0. The image results demonstrate the blurring effect of edges present on both mean and Gaussian filtered results. This effect can be clearly seen from the edge map images. Median filters on the other hand can be observed to retain edge information properly while effectively removing noise in the image. But, median filters have the tendency to shift the image boundary information when used with a large neighborhood [57]. Therefore for this research, a 3x3 neighborhood was selected for the 2D median filter implementation.

Active contours or snakes have been utilized in image segmentation for its ability to converge towards the closest local minima from its initialized position in an iterative process which moves each of its control points towards image gradients and therefore yielding a better approximation for the object outline. One of the most profound limitations of the physics based snake or active contours is the inability



(a)Original image

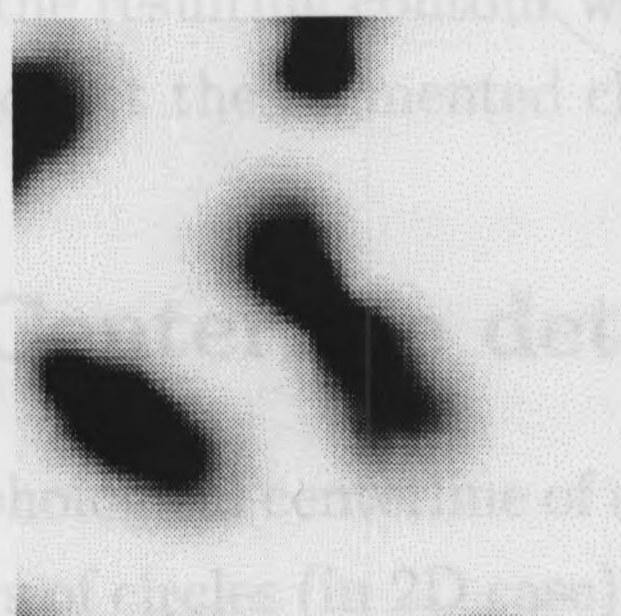
(b)Original edge map



(c)Median filtered image



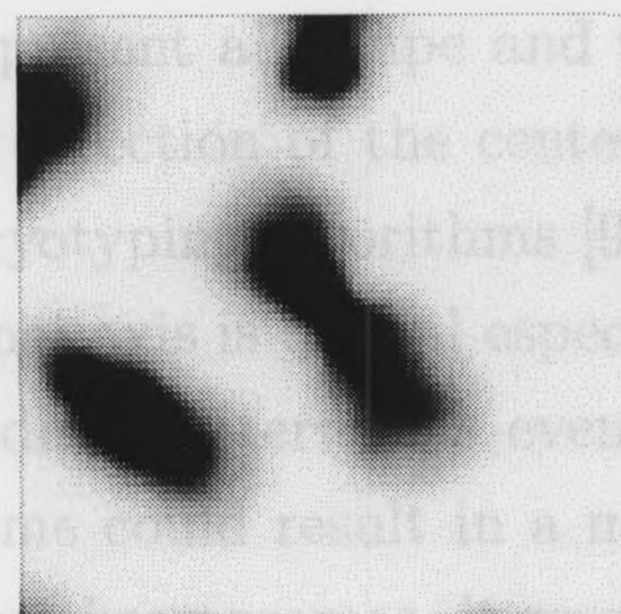
(d)Median filtered edge map



(e)Mean filtered image



(f)Mean filtered edge map



(g)Gaussian filtered image



(h)Gaussian filtered edge map

Figure 3.6: Application of different image filters (mean, Gaussian and median) and their corresponding edge maps. All the filters were applied to a neighborhood of 5×5 region.

to observe the image gradients which are beyond the current control point positioning. Images of human metaphase chromosomes have a high tendency to bend readily during sample preparation process and thus create concave boundaries which a conventional snake would not converge towards. Smoothing the edge map of an image with a Gaussian kernel is one of the most commonly used methods to increase the capture range. This provides very little range improvement and also results in inaccurate object segmentation due to the washed out boundaries by the gaussian kernel. Out of many other attempts that have been taken to address this problem such as by using distance potential forces [37],[38], the Gradient Vector Field (GVF) approach seems to provide better capture range for snakes [28]. External energy models for increasing the capture range, along with the advantages of using GVF snakes are discussed in section 2.3. The "Canny edge detection operator" which uses a multi-stage algorithm to accurately detect image boundary edges was utilized for generating the edge map [22]. We selected GVF snakes because of these advantages. Based on our experience, an iteration count of 125 was used to achieve reasonable results. Control points of the resulting contour were used to create a binary mask which in return was used to extract the segmented chromosome from its background.

3.2 Centerline detection

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 [58]. The centerline of a chromosome should ideally represent all shape and topological information of the original chromosome. Accurate detection of the centerline of chromosomes is a critical operation in most of the karyotyping algorithms [9],[10],[11],[13]. The accuracy of the centerline or the longitudinal axis is crucial especially when classifying the chromosomes according to their banding patterns, as even a small deviation of the centerline away from the chromosome could result in a misclassification. Yet, due to the shape variability of metaphase chromosomes, it remains difficult to extract the centerline properly from chromosomes.

A detailed literature survey on existing centerline extraction methods is provided in section 2.2.2. The majority of chromosome centerline finding methods are performed with the following two steps,

1. Iterative morphological thinning or skeletonization process which reduces the 2D binary image to a collection of points representing the original shape information of the object.
2. A pruning process which is used to remove all the spurious branches present in the skeleton obtained above.

While thinning algorithms tend to perform well, the second step (pruning) is found to be a harder problem to address. Skeletonization process is overly sensitive to boundary deformations, noise and irregularities and that proves to be a significant drawback in an application point of view. Skeletonization methods provide a structure with numerous small branches which can cause difficulties for any higher level image or shape analysis. Also, when dealing with digital images, image boundaries are bound to be discontinuous due to quantization errors, irrespective of the resolution. Skeletonization as well as thinning methods try to preserve these discontinuities in the skeleton and therefore, increase the probability of getting spurious branches. Figure 3.7 shows some skeletonization and thinning results [5] of chromosomes of different shapes. The unpredictable branches and the presence of other unwanted information in the skeleton are the issues with commonly used methodologies in which the algorithms attempt to preserve all boundary information for re-construction of the object (as seen in figure 3.7). Though thinning yields less spurious branches compared to skeletonization (refer figure 3.7), it could still provide unwanted branches specially near the telomere regions and also at boundary concavities. Therefore all these factors pointed out the necessity of proper pruning in order to obtain an accurate and a reliable centerline.

Skeleton pruning can be done either during the iterative process of building the skeleton itself or as a post processing step in which the pruning methods are applied after obtaining the skeleton of the structure [59]. Many of the existing work done on pruning are "application oriented" and are based on simple methodologies. The most common pruning method is based on the Prairie fire model [58], in which the propagation speeds/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. Thresholding of the skeletal points based on the shape contribution of each point was another commonly adopted method. But, without proper regularization, this also has shown to provide disconnected skeletal segments.

Observing these limitations with respect to the demands in our specific application, the following were laid out as the target characteristics of a proper skeleton pruning method.

- The positioning of the skeleton must be accurate and should not deviate due to pruning (contain exact maximal disk centers and thus would be the symmetrical axis).



(a)Original image



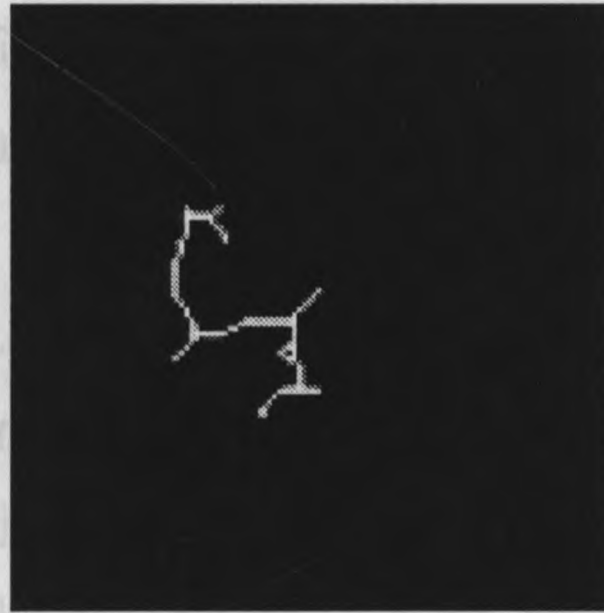
(b)Skeleton result



(c)Thinned result



(d)Original image



(e)Skeleton result



(f)Thinned result

Figure 3.7: Two chromosomes and their skeletonization [4] and morphological thinning results [5] showing some resulting spurious branches. These operations were performed on the binary object obtained through Otsu's method.

Then, provided that we have the following,

- A partitioning Γ of the object boundary ∂D
- The skeleton of the original object, denoted by $S(D)$.

Then, its possible to find the generating points $Gp(s)$ using,

Observing these limitations with respect to the demands in our specific application, the following were laid out as the target characteristics of a proper skeleton pruning method.

- The positioning of the skeleton must be accurate and should not deviate due to pruning (contain exact maximal disk centers and thus would be the symmetrical axis).
- All the shape features must be retained in the skeleton as much as possible.
- The resulting skeleton should be consistent with a wide range of shape variability and object orientation.

We have adopted a skeleton pruning method based on Discrete Curve Evolution (DCE) [42] that achieves all of the above requirements. The DCE based pruning in our algorithm was applied only to chromosomes with skeletons longer than a particular length (35 skeletal points) and shorter chromosomes were processed using the thinning algorithm described by Lam [5]. The rationale behind this hybrid application is to use DCE based pruning only on chromosomes which are highly likely to be bent while utilizing thinning on relatively shorter chromosomes for which the skeleton deviates from the centerline. This skeleton pruning process was based on first partitioning the object contour into few polygonal sections and then pruning by removing all skeletal points of which all the generating points (the points where the maximal disks touch the object boundary at more than 1 point) lie on the same polygon partition.

In order to provide a formal definition for this DCE based pruning method [42], let D be a planar set which encloses a connected bounded open subset in \mathbb{R}^2 and ∂D be the boundary of that planar set. For this method, the D set must be simply connected. In other words, the object must not be twisted or contain any holes. Then, provided that we have the following,

- A partitioning Γ of the object boundary ∂D
- The skeleton of the original object, denoted by $S(D)$.

Then, its possible to find the generating points $Gp(s)$ using,

$$Gp(s) = \{\forall s \in S(D) \mid p \in \partial D \text{ and } p \in B(s)\}$$

where p is a set of points with $|p| \geq 2$ and the $|\cdot|$ operator been the cardinality operator. Also, $B(s)$ is the points on the circumference of the maximal disk with its center at point s of the skeleton ($s \in S(D)$). This is the largest disk that is completely contained within the object and not a subset of any disk with the same center. Therefore, the generating points given by Gp is a subset of ∂D . Then, the skeleton pruning process can be defined as the removal of all points $s \in S(D)$ which has its generating points ($Gp(s)$) on the same open segment (Γ^i) of the partitioning Γ [42]. Therefore, this only leaves skeletal points that have generating points ($Gp(s)$) lying on different open contour partitions Γ^i & Γ^j where $i \neq j$.

Results obtained with the above skeletal pruning method are highly dependent on the contour partitioning (Γ) used for the algorithm. Some partitions suit better than the others and the best partitions are observed to be those which represent high level shape contribution of the original object. On top of that, to acquire a pruned skeleton with lesser branches, a partitioning $\Gamma = \{\Gamma^0, \Gamma^1 \dots \Gamma^n\}$ is needed with a low value for n . Therefore the skeleton pruning problem can be viewed as a contour partitioning and a shape simplification problem. DCE provides an ideal solution for this by effectively evolving polygon partitions by vertex deletion based on any given relevance measurement [43]. Digital image boundaries can be approximated to a polygon without a loss of information by taking each boundary pixel as a vertex on the polygon and similarly considering the distance between each pixel as edges. DCE has the capability to evolve the polygon iteratively by removing the vertex which has the least value for the defined relevance function given by equation 3.4. By carefully selecting a proper function to calculate the relevance value in a way that becomes dependent on features of its neighbors, DCE can evolve using global features of the shape rather than local information. Also, as DCE is simply deleting vertices of the polygon partitions, the topological information is guaranteed to be represented in the resulting skeleton. Furthermore, in our research, only the convex polygon combinations were considered in order to effectively prune spurious branches [3],[42].

The relevance function $K(v, u, w)$ [42],[43] at point v with points u & w as its neighbors is given below in Eq 3.4,

$$K(v, u, w) = (\theta * d_{uv} * d_{vw}) / (d_{uv} + d_{vw}) \quad (3.4)$$

where d_{uv} & d_{vw} are the Euclidian distance between the vertices (normalized with respect to the length of the contour) and θ is the turn angle at vertex v . A detailed description regarding the DCE algorithm along with the rationale behind using the given relevance function (in equation 3.4) is provided in section 2.4.

When configured as described above, Discrete Curve Evolution delivers the following advantages to the pruning process[42],[43],

- As DCE simply deletes vertices without moving them, there is no dislocation of feature points and effectively no dislocation of the skeleton as well.
- Can handle complex shapes. Provision has to be taken to avoid self intersecting.
- Effectively removes noise from the setup and thus eliminates the effects of digitization noise and any other source of noise present in the digital image without effecting boundary information.
- The convergence is guaranteed (based in global contour information) and stopping criteria is simple to set.

In the case of obtaining the centerline of a chromosome, the ideal case would be to obtain a pruned skeleton with no extra branches. Yet, as the minimum non-trivial polygon being a triangle and DCE being modeled as polygons, the resulting skeleton will at minimum have one spurious branch. Although this is not the ultimate result required for sampling the centerline, it still provides a very reliable method to have a known number of branches (3 branches in this case) for a chromosome of any shape and orientation. Thus, the use of this method eliminates the unpredictability of the conventional thinning or skeletonization (medial axis transform) methods and provides a good starting point at getting the actual longitudinal axis of the metaphase chromosome. Figure 3.8 depicts the above mentioned reliability that can be achieved through DCE based pruning when compared to standard skeletonization.

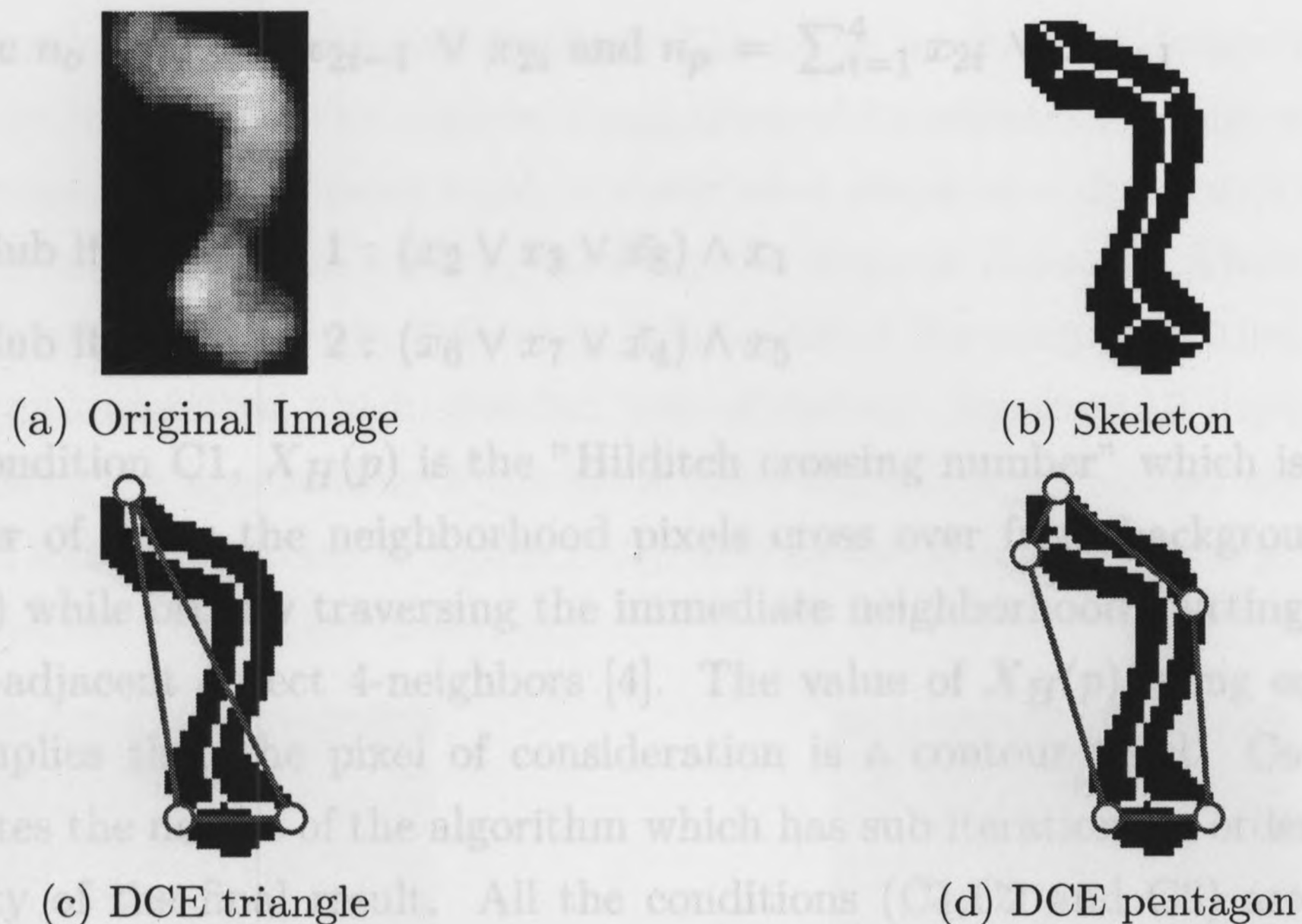


Figure 3.8: Comparison between standard skeleton with DCE based solutions where the figure 3.8(b) is the skeletonization result. The marked points figure 3.8(c) & figure 3.8(d) depict the end polygon vertices of the DCE process with different end conditions.

Though most of the regular thinning or skeletonizing algorithms provide a connected one-pixel thick skeleton [31], it is not guaranteed in this approach. Nevertheless, we opted for a reliable shape representation with a known number of branches present in the result. Hence, the skeleton achieved through the DCE based pruning method was further processed using a modified thinning operation. The basic thinning algorithm [5] used was a parallel thinning algorithm in which the deletion of a pixel was performed "if and only if" all the following three (3) conditions (C1, C2 and C3) were satisfied,

- C1 : $X_H(p) = 1$

in which,

$$X_H(p) = \sum_{i=1}^4 b_i$$

$$b_i = \begin{cases} 1 & \text{if } x_{2i-1} = 0 \text{ and } (x_{2i} = 1 \text{ or } x_{2i+1} = 1) \\ 0 & \text{otherwise} \end{cases}$$

- C2 : $2 \leq \min\{n_o(p), n_e(p)\} \leq 3$

where $n_o = \sum_{i=1}^4 x_{2i-1} \vee x_{2i}$ and $n_p = \sum_{i=1}^4 x_{2i} \vee x_{2i+1}$

- C3 :

Sub iteration # 1 : $(x_2 \vee x_3 \vee \bar{x}_8) \wedge x_1$

Sub iteration # 2 : $(x_6 \vee x_7 \vee \bar{x}_4) \wedge x_5$

In condition C1, $X_H(p)$ is the "Hilditch crossing number" which is defined as the number of times the neighborhood pixels cross over from background ('0') to object ('1') while orderly traversing the immediate neighborhood, cutting the corner between 8-adjacent object 4-neighbors [4]. The value of $X_H(p)$ being equal to one directly implies that the pixel of consideration is a contour pixel. Condition C3 demonstrates the nature of the algorithm which has sub iterations in order to enforce connectivity of the final result. All the conditions (C1, C2 and C3) are meant for identifying the connectivity of the pixel of interest with its neighbors in order to decide whether the removal of the pixel would effect the connectivity of the structure or not. For all the above conditions, the (3x3) neighborhood is organized and referred as depicted in figure 3.9 in which $x_9 = x_1$.

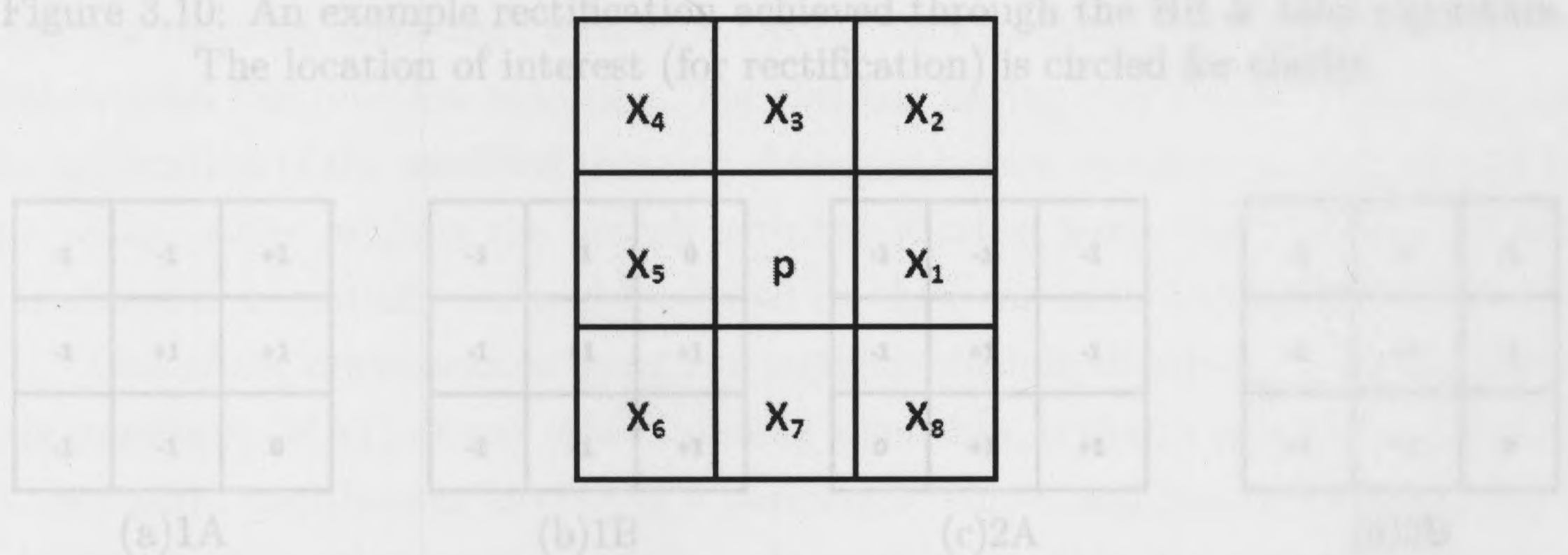


Figure 3.9: The 3x3 neighborhood setup for the thinning process where p is the pixel of interest and $X_9 = X_1$.

A modification for the above thinning algorithm was required to cater to a specific concern that arose as an effect of the nature of the skeleton obtained through the DCE based pruning method. A new step was introduced as a pre-process to the thinning operation which intended to rectify these specific instances. This was achieved by exposing the DCE skeleton to a series of masks, based on the "morphological hit & miss" operation and then subtracting this result from the binary skeleton image

obtained through the DCE based method. The set of masks were designed to specifically identify the pixels to be removed regardless of its orientation. This was achieved by taking rotations of the basic mask to cover all 4 major axes degrees (8 masks) and the result of this stage was subtracted from the original skeleton. These masks are given in figure 3.11. Thus, by following this modified thinning algorithm, a reliable, connected and one-pixel thick skeleton was obtained. Figure 3.10 depicts such an instance where the application of the hit & miss algorithm mentioned above rectifies the marked location (circled in figure 3.10), and by doing so allows the pruning stage (described next) to function effectively.

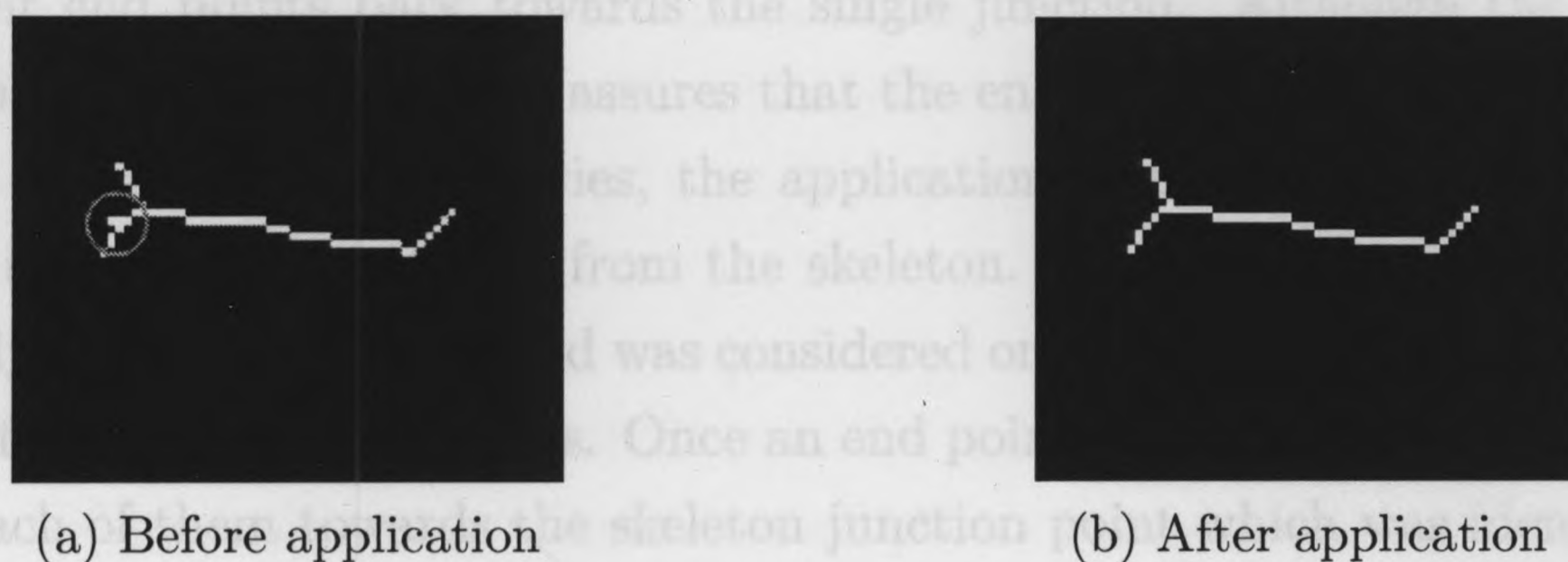


Figure 3.10: An example rectification achieved through the Hit & Miss algorithm. The location of interest (for rectification) is circled for clarity.

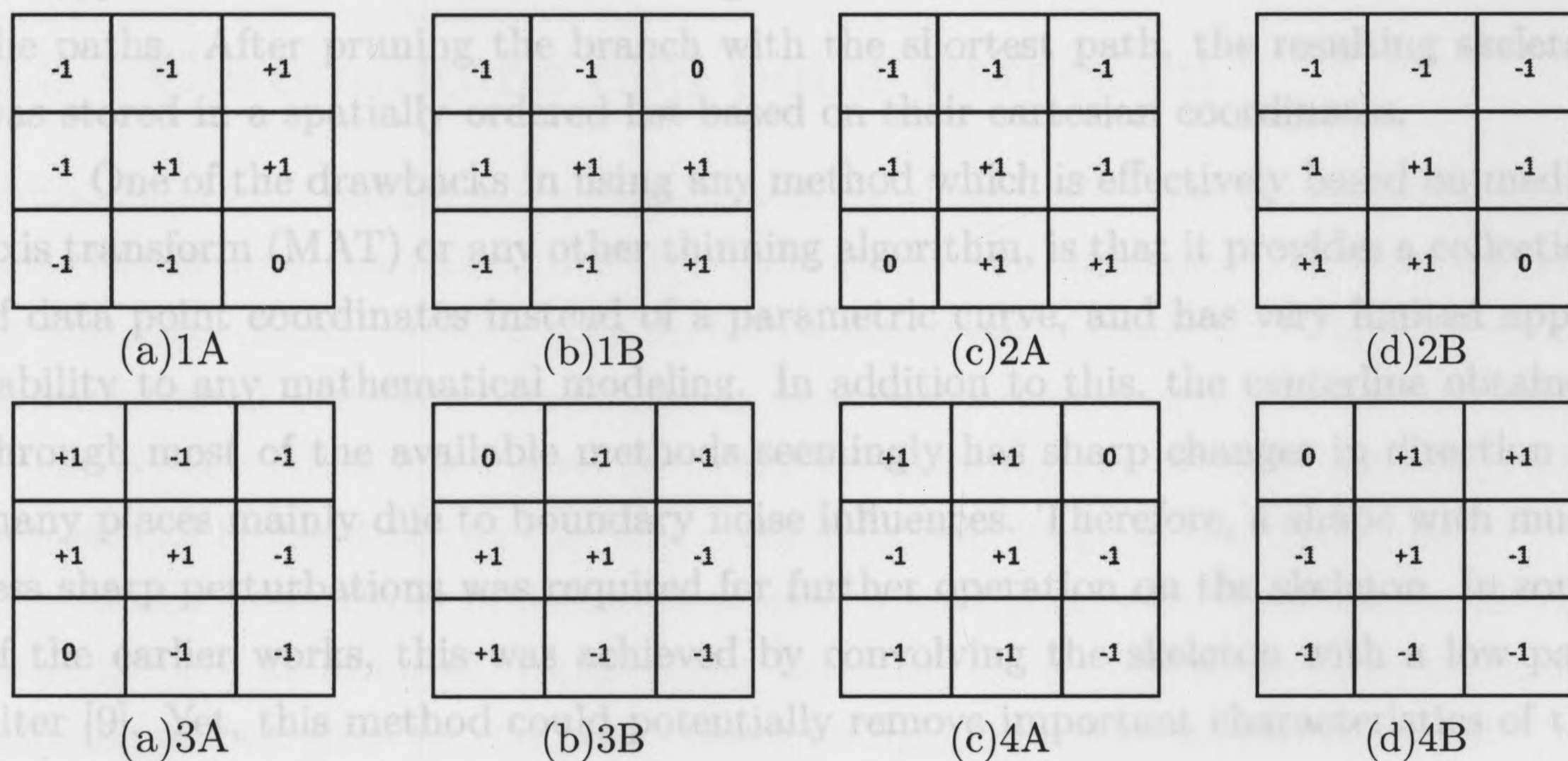


Figure 3.11: The masks used for the Hit & Miss process where '-1' as -1, '+1' as 1 and '0' as ignored.

The skeleton achieved through this process contains a total of three branches. In order to accomplish a skeleton with no spurious branches, one of these must be pruned while retaining shape features of the object within the skeleton. Based on the observations of skeletal results (based on the DCE setup) of different chromosomes over a variety of chromosomal shapes, the shortest skeletal branch was selected to be pruned. As the DCE based pruning method is based on the topological skeleton, this branching occurs either at one end of the chromosome or at one of the bending locations. Therefore, this observation supported the decision to prune the shortest branch. Therefore in this particular case, all the skeletal branches needed to be traced from their end points back towards the single junction. Although the DCE based approach into skeleton pruning assures that the end points coincide with the original sampled digital image boundaries, the application of the thinning step potentially removes some end data points from the skeleton. Thus, to find the end points, an iteratively growing neighborhood was considered on the image using the DCE skeletal end points as the starting points. Once an end point was obtained, the program then traced each of them towards the skeleton junction point which was identified by the characteristic of having multiple responses for the tracing filter. The above described modified thinning algorithm assures a 1-pixel thick skeleton with only one position that satisfies this junction condition. The skeleton having only three (3) branches and the application of the modified thinning discussed before, assured accurate results for the paths. After pruning the branch with the shortest path, the resulting skeleton was stored in a spatially ordered list based on their cartesian coordinates.

One of the drawbacks in using any method which is effectively based on medial axis transform (MAT) or any other thinning algorithm, is that it provides a collection of data point coordinates instead of a parametric curve, and has very limited applicability to any mathematical modeling. In addition to this, the centerline obtained through most of the available methods seemingly has sharp changes in direction at many places mainly due to boundary noise influences. Therefore, a shape with much less sharp perturbations was required for further operation on the skeleton. In some of the earlier works, this was achieved by convolving the skeleton with a low pass filter [9]. Yet, this method could potentially remove important characteristics of the centerline especially when subjected to a centerline of a bent chromosome. Therefore a curve fitting step was used to obtain this smooth curve from the skeleton of the previous step. Few methods including polynomial approximation were explored in

order to find a method that yields an accurate centerline approximation. Finally, a cubic spline interpolation based method was selected, which attempts to fit a 3rd order polynomial between each of its control points (knots) while keeping continuity at its end point connections. A detailed theoretical analysis of spline curve fitting is given in section 2.5. Yet, the DCE resulting skeleton ends need to be pruned before fitting a spline. This is due to effects introduced by the bifurcations of the initial skeleton (prior to pruning of the extra branch) near telomere regions of the chromosome. Therefore, the centerline was pruned at the ends by 10% (empirically set) of the total length of the centerline. Then, the control points for curve fitting were provided by sampling this pruned skeleton result and registering a control point for approximately every 7 skeleton points. The interval of '7' points above was selected empirically in order to avoid over fitting the data while representing the shape information adequately.

Cubic spline interpolation was performed with respect to both x and y axes in order to assure a proper distribution in the spline interpolated result as the DCE based centerline could potentially have multiple x or y coordinates. Then, these were combined and duplicates were detected and removed. The ends of the centerline were clipped to remove the skeletal portion that deviates at the telomere regions from the actual centerline.

3.3 Chromosome end point detection

Methods that use medial axis transform tend to either yield bifurcations or deviations from the desired centerline, near the ends of the chromosome. Thinning methods on the other hand, provide a centerline which doesn't reach the boundary of the chromosome at the ends. Therefore, in both these scenarios, a method is required to correct or extend the centerline to the telomere regions of a chromosome. This process can be identified as the 'end point detection' stage of the algorithm.

Methods available in literature related to end point correction are simple and heuristic. For an example, Wang simply extended the centerline obtained through a thinning process from its end points based on the gradient at the last two points [32]. Therefore, we have also resorted to a heuristic approach for this purpose. In our research, we used a gradient (edge) and intensity based end point detection method. In order to achieve this goal we used the same set of sample points that were used

earlier for spline curve fitting step as this choice of points improved the results by a significant margin. Then a template was created in order to match and detect intensity information specifically present at the end points of a chromosome (see table 3.2). It consisted of three (3) regions namely background (-1), object (+1) and ignore (0), which had different contributions in polarity towards the image features. Next, this mask was rotated to the proper angle using the standard rotation matrix and nearest neighbor interpolation, where it was then convolved with the image after being properly padded with zero (0) values at the boundaries. The convolution was done on the line created by extending the two sample points at the end and was incremented based on either the 'x' axis or the 'y' axis which was decided based on the gradient of this line. The lengths of these segments were selected to be 20% of the centerline.

The scores at each location along a line extended from the end points were recorded and the maximum response and its index was selected. Next, the score values were examined for their gradient along the line from the maximum response point by sampling intensity values along the extended line segment. The end point was selected by applying a threshold for this gradient value and thus assuring a consistent end point while accounting for intensity fading in the fluorescence microscopic images.

Table 3.2: A 7x7 representation of the original 20x20 template used for end point correction, where the coefficients were set as '0' - ignored, '+' as +1 and '-' as -1.

0	0	0	0	0	0	0
0	0	-	-	-	0	0
0	0	-	-	-	0	0
0	-	-	+	-	-	0
0	-	+	+	+	-	0
-	+	+	0	+	+	-
0	+	0	0	0	+	0

3.4 Centromere identification and polarity assignment

The centromere is the most condensed and constricted region of a chromosome, to which the spindle fiber is attached during mitosis (cell division) [32]. It is also the location where the two sister chromatids join together. The detection of centromere is of utmost importance in most karyotype classification methods as the class of the chromosome is dependent upon the location of the centromere with respect to the chromosome end points. The centromere index is the ratio between the short arm to the total length of a chromosome, which corresponds to the location of the centromere. The centromere of a chromosome is characterized by its constriction near the centromere region (refer section 1.1.1.1 & section 1.1.1.3 for more details). Yet, when dealing with real chromosomal images, an untrained human eye could find it hard to accurately detect this constriction, especially on bent chromosomes. Quantization noise and image noise could add on to this effect and therefore makes the centromere detection even harder. The difficulty increases as the banding number of the chromosome image (which corresponds to the length of chromosomes) increases. Also, the constrictions could vary depending on the type of the chromosome.

The constriction is more pronounced in metacentric chromosomes where the centromere is located more towards the mid point of the structure and less pronounced in acrocentric chromosomes which have the centromere towards an end point of the chromosome. This simple location difference and their appearance severely limits the detection of the centromere in acrocentric chromosomes. There have been many research work carried out in the field of karyotyping, where the centromere location information was used to classify chromosomes into different classes [9],[18],[19],[32],[60]. Some of the centromere detection algorithms were based on methods that did not involve finding the centerline of the chromosome. Mousavi assigned a membership value for each pixel of DAPI and FITC (Fluorescein Isothiocyanate) images (with centromere probes) based on an iterative fuzzy algorithm [60]. Yet, this method has limited scope of application as it depends on special specimen preparation and information in the form of centromere probes and FITC images. Another work carried out by Moradi [18] and similarly by Faria [61] (on fish chromosomes) took the horizontal and vertical projection vectors of the binary segmented chromosomes. These projection vectors were obtained by summing up the number of object pixels in the binary

segmented image in each horizontal and vertical directions and the centromere was located by finding the global minimum in these vectors. This method did not perform satisfactorily for both acrocentric chromosomes as well as for any chromosome with a bend greater than 90^0 degrees.

Majority of work carried out on centromere location is based on getting the centerline of the chromosome first. Piper's approach towards this was to achieve the second moment of the profile of the chromosome along its centerline [9]. Yet, this method depends on the accuracy of the centerline achieved using the medial axis transform and "poor man's skeleton" which was not reliable (see section 2.2.2.1). Moradi took the average of image intensities along scan lines perpendicular to the centerline and used wavelet de-noising to remove sharp perturbation in the density profile (DP) [19]. He then classified the chromosomes with the use of a trained artificial neural network (ANN). He also relied on medial axis transform for the centerline and his method could only prune branches that occurred at the ends of the chromosomes. Therefore, this method would yield unsatisfactory results when branching occurs away from the chromosome ends. On top of that, uniform averaging of intensities on the trellis would make the sampled values highly susceptible to noise especially at bent regions of chromosomes. Wang extracted the shape profile, density profile and the banding patterns using scan line sampling and then used a rule based setup to detect the chromosome centromere which claimed to have improved the reliability of the result [32]. His approach as well as many other karyotype analysis methods are developed for Geimsa banded (G-banded) chromosomes unlike the DAPI stained chromosome images which are used in our research. Geimsa banded chromosomes are known to have clearer visible chromosomal banding patterns. On top of that, Wang also relied on thinning to get the centerline, which is also prone to spurious branching as discussed in section 3.2.

Therefore, in our approach, we have used the same end-pruned centerline which was used for getting sample points in section 3.3 (spline curve fitting) as the reference. This selection of the reference line segment, drastically simplified the centromere detection process as this excludes the extreme ends of the chromosome. Then, line segments (referred to as 'trellis' in subsequent sections) were drawn perpendicular to that of this 'pruned centerline' segment at unit length intervals. Figure 3.12 illustrates an example of a trellis structure and the GVF result outline (superimposed) which is used for the centromere detection stage. The sample intensities along the trellis

Therefore, having this information, the challenge was in developing a suitable framework for combining the data features to get the centromere location. We designed a simple methodology based on the concepts of 'model based segmentation' in which features were combined using a prior (scaling factors) which bias the suitable according to the instance (chromosome type) over the other. Metacentric chromosomes are known to have a clear and more pronounced centromere whereas acrocentric chromosomes are not. On top of all these, in this work we had a prior knowledge of the chromosome which we were

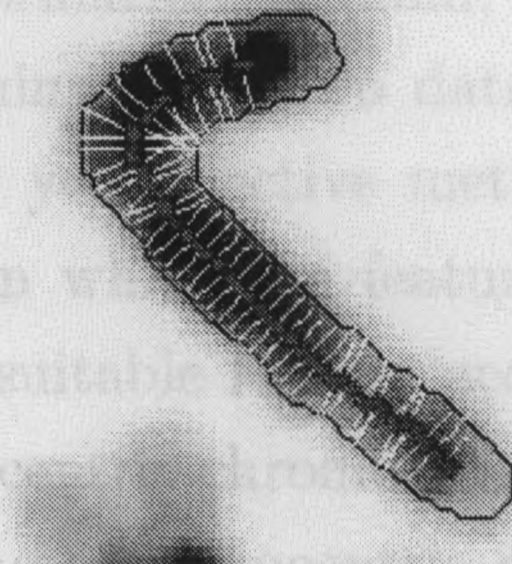


Figure 3.12: The trellis (in yellow) and the superimposed GVF contour (in blue) used for the centromere identification.

factors were calculated. features were appropriately combined into a single feature set. Centromere location was then obtained were weighted with a Gaussian function which was intended to cancel image and boundary noise as well as effects introduced by bending of the chromosomes. Though the sampling of intensities along the trellis was performed on the filtered DAPI image, the length of the trellis segments were decided from the binary result obtained through GVF in section 3.1. The motive behind using the GVF result was to base the trellis on a binary image which had more edge characteristics than a simple thresholded binary image making the constriction at the centromere more pronounced. Therefore in our approach, we relied on the following parameters to locate the centromere,

3.4.1. The centromere confidence measure

1. The width profile of the chromosome along the trellis on the centerline, which was obtained using the GVF binary image result.
2. Density Profile (DP) obtained by getting the weighted average of intensity values of the DAPI image (based on a Gaussian function) along the trellis limited by the GVF result.

All the previously mentioned approaches in detecting the centromere are mainly limited by the lack of knowledge of the information relevant to chromosomes as these are merely a section of a karyotype analysis problem. Our application differed from karyotype analysis as the system was meant to analyze a known chromosome at any given instance. Therefore, the centromere identification process can be adapted to include that information to assist the detection process. For example, the information whether a given chromosome is acrocentric or not would greatly aid the process of finding the centromere.

Therefore, having this information at hand, the challenge was in developing a suitable framework for combining these two data features to get the centromere location. We designed a simple yet, effective methodology based on the concepts of 'model based segmentation' in which the features were combined using a *prior* (scaling factors) which bias the suitable feature according to the instance (chromosome type) over the other. Metacentric chromosomes are known to have a clear and more pronounced centromere whereas acrocentric chromosomes are not. On top of all these, in this work we had a prior knowledge of the chromosome which we were dealing with. Thus, based on the chromosome number, probability values or biasing factors were calculated using the "biasing prior" by which the features were appropriately combined into a single feature set. Centromere location was then obtained by finding the global minimum in this resulting feature set.

After obtaining the centromere location as described above, the polarity can be easily assigned using the knowledge about chromosome groups (refer 1.1.1.1). So, as we have the telomere locations and the centromere location along the centerline, the Euclidian distances to the centromere from each telomere can be easily calculated. Finally, by simply comparing these distances, the short arm (p-arm) and the long arm (q-arm) can be identified and therefore the polarity can be assigned.

3.4.1 The centromere confidence measure

The centromere detection methods in literature simply attempt to detect the centromere location of a chromosome. Yet, accurate detection of this location is a difficult process which demands the aid of a well trained expert. Therefore, all the existing methods (including the method discussed in this thesis) are prone to some errors in detection. In this research, we propose a measure termed as the 'centromere confidence value' or CCF which will yield the level of probable accuracy of an automated detection process. The objective here is to provide a percentage value (0% - 100%) to the expert, which gives an approximate level of confidence of the detected location. A proper confidence value could directly assist cytogeneticists in their decision making process. Thus we could overcome an inherent problem in the automated detection process.

Majority of methods that detects the chromosome centromere are based on the chromosome centerline. These methods in general, depends on the following three

features for the centromere detection.

1. The width profile
2. The density / intensity profile
3. The banding patterns

Yet, the banding pattern feature is predominantly used for G - banded chromosome analysis. This is because other staining methods such as 'DAPI' does not yield a clear banding pattern as with Geimsa (G) staining. Therefore, the width and the intensity profiles can be considered as the two most practical measures. Thus our proposed Centromere confidence (CCF) measure is defined based on those two feature profiles. Lets define $Wp(i)$ and $Ip(i)$ respectively as the width and intensity profiles of centerline points $i = 1, 2, \dots, N$. Also let Sp be the global minimum (selected as centromere) of the combined profile (discussed in section 3.4). Then, the values W_m and I_m can be defined as follows,

$$W_m = \min_i \{Wp(i)\}$$

$$I_m = \min_i \{Ip(i)\}$$

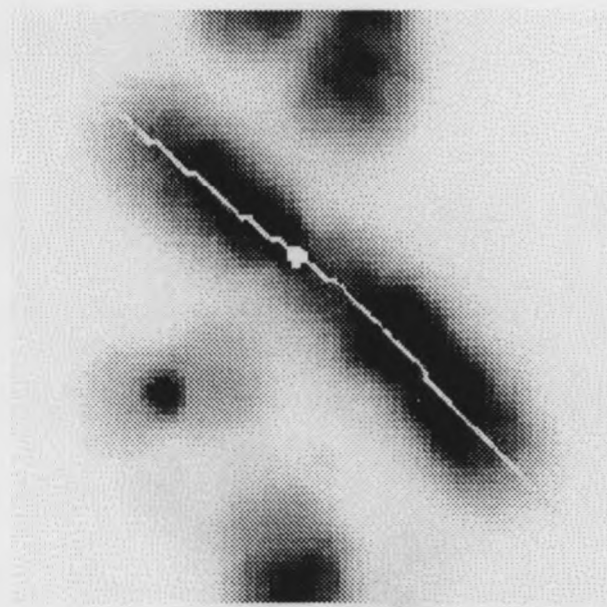
Next, the vector V_{wp} that contains all the centerline points with the minimum width value can be defined as follows,

$$V_{wp} = \{u \in Wp(i) \mid Wp(i) = W_m, \forall i = 1, 2, \dots, N\}$$

Then, the value Wid_{Wp} is calculated as the cardinality of the above vector V_{wp} ($Wid_{Wp} = |V_{wp}|$). Therefore, the centromere confidence measure 'CCF' is defined as given by equation 3.5.

$$\text{CCF} = \frac{(W_m + I_m)}{(Sp \times F_c)} \times 100\% \quad \text{where } F_c = \begin{cases} (.55 \times Wid_{Wp}), & \text{if } Wid_{Wp} > 1 \\ 1, & \text{if } Wid_{Wp} = 1 \end{cases} \quad (3.5)$$

The rationale behind the 'centromere confidence value' (CCF) can be explained with the aid of figure 3.13 and figure 3.14. The chromosome analyzed in figure 3.13(a)



(a) Chromosome with the centerline (green) and the detected centromere (white) marked.

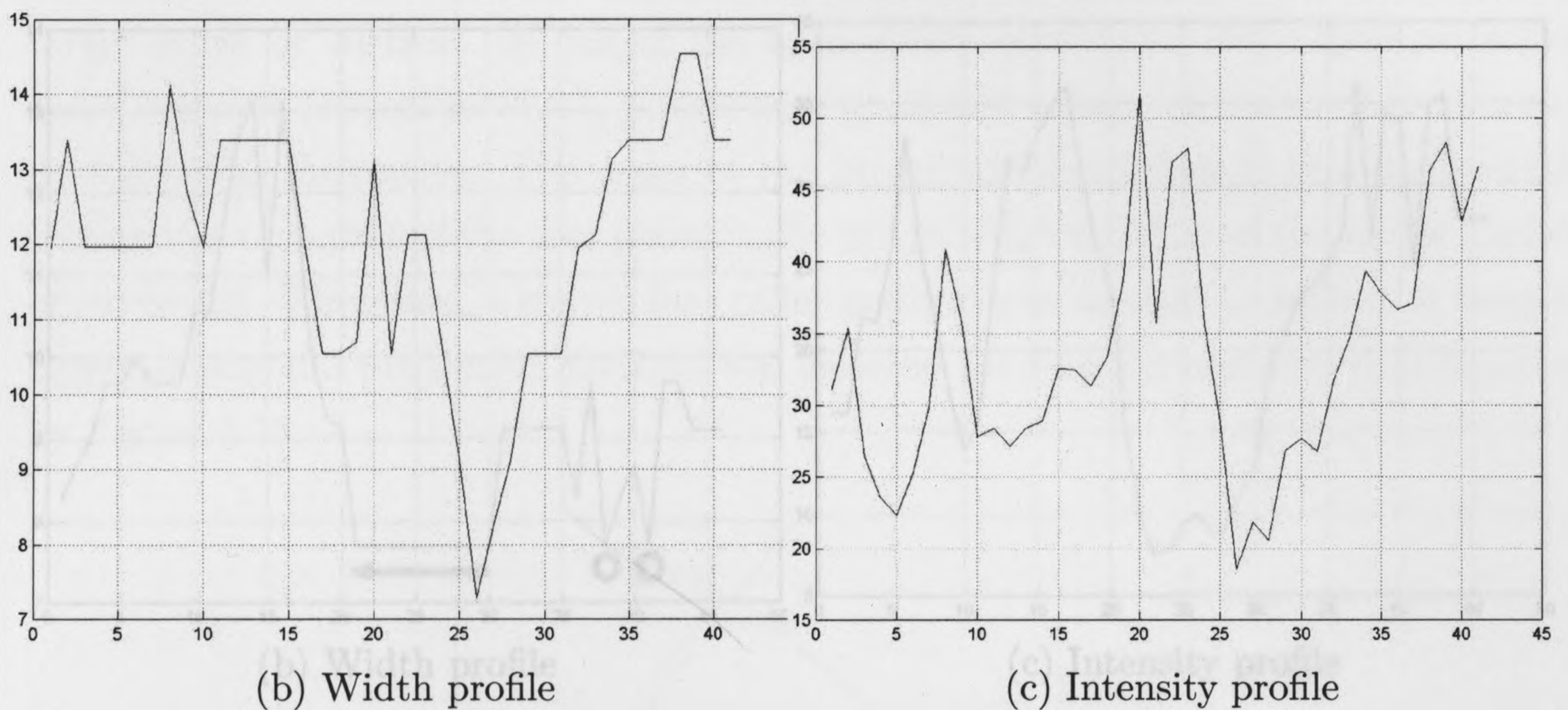


Figure 3.13: The width profile (figure 3.13(b)) and the intensity profile (figure 3.13(c)) of a straight chromosome which is shown in figure 3.13(a). The x axis in figure 3.13(b) & figure 3.13(c) is the sample data points of the centerline (from bottom to top of the chromosome) while the y axis are respectively the width and the image intensity values at that point.

is a straight chromosome and the corresponding intensity and width profiles have a global minimum at the 26th data sample point. The clear intensity change near the centromere region in figure 3.13(c) can be considered as a special case. Yet, it is the sharp and unique change in the width profile (figure 3.13(b)) that gives a concise evidence of the constriction. Figure 3.14(a) depicts a bent chromosome, where the detection of the centromere becomes more difficult. The width profile given by figure 3.14(b) demonstrates multiple points with the minimum width value. This is mainly caused by the overlapping of the trellis section (discussed in section 3.4) due to the bends on the chromosome. Intuitively, in such a scenario, a lower confidence level (in the centromere detection) should be presented.

3.5 FISH probe projection

In order to obtain the fractional ratio of the FISH probe locations, the accurate and reliable projection of the probe signal onto the measuring grid (which is the centerline in this case), is of utmost importance. Various point projection methods were developed over the years which are based on the second order ('order' defines the derivative used) [62],[63],[64]. In the literature a lot of methods were proposed such

(a) Chromosome with the centerline (green) and the detected centromere (white) marked.

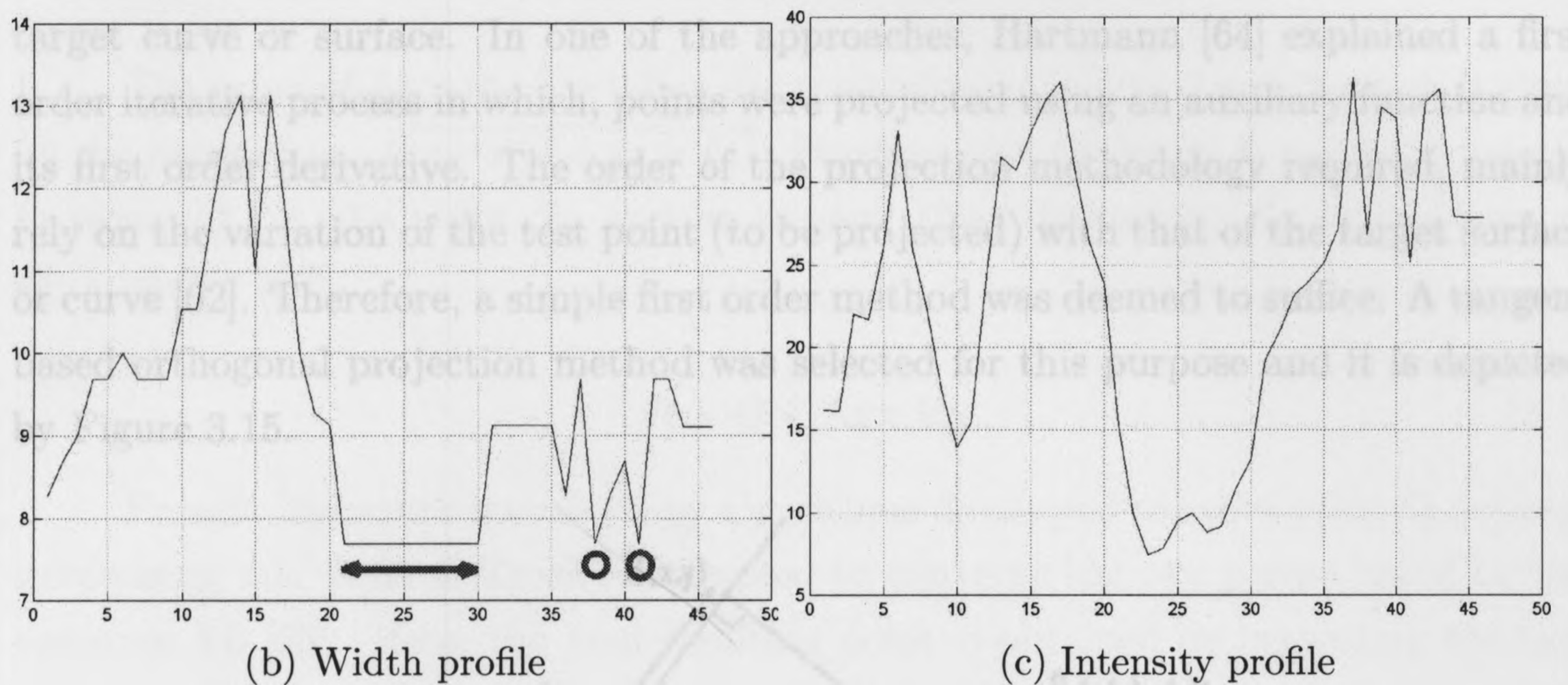


Figure 3.14: The width profile (figure 3.14(b)) and the intensity profile (figure 3.14(c)) of a 'bent' chromosome which is shown in figure 3.14(a). The x axis in figure 3.14(b) & figure 3.14(c) is the sample data points of the centerline (from top to bottom of the chromosome) while the y axis are respectively the width and the image intensity values at that point. The points which corresponds to the minimum width profile value is marked in figure 3.14(b).

The CCF value given by equation 3.5 yields an high value if the added minimum of the width and intensity profiles are equal or closer to the selected combined minimum point (centromere). Also, there is the possibility of getting a width profile with multiple points with the global minimum value ($Wid_{Wp} > 1$). This occurs as the trellis segments can miss the actual constriction either due to high bends or small perturbations in the centerline of the chromosome. The variable ' F_c ' is defined to specially reduce the confidence value of the CCF value in the presence of such multiple ($Wid_{Wp} > 1$) minimum points (refer figure 3.14(b)).

3.5 FISH probe projection

In order to obtain the fractional ratios of the FISH probe locations, the accurate and reliable projection of the probe signal onto the measuring grid (which is the centerline in this case), is of utmost importance. Many iterative point projection methods were developed over the years which are of zero, first and second order ('order' defines the derivative used) [62],[63],[64]. In the literature, a lot of methods were proposed such as Iterative Closest Point (ICP) which attempted to map a given set of points to a target curve or surface. In one of the approaches, Hartmann [64] explained a first order iterative process in which, points were projected using an auxiliary function and its first order derivative. The order of the projection methodology required, mainly rely on the variation of the test point (to be projected) with that of the target surface or curve [62]. Therefore, a simple first order method was deemed to suffice. A tangent based orthogonal projection method was selected for this purpose and it is depicted by Figure 3.15.

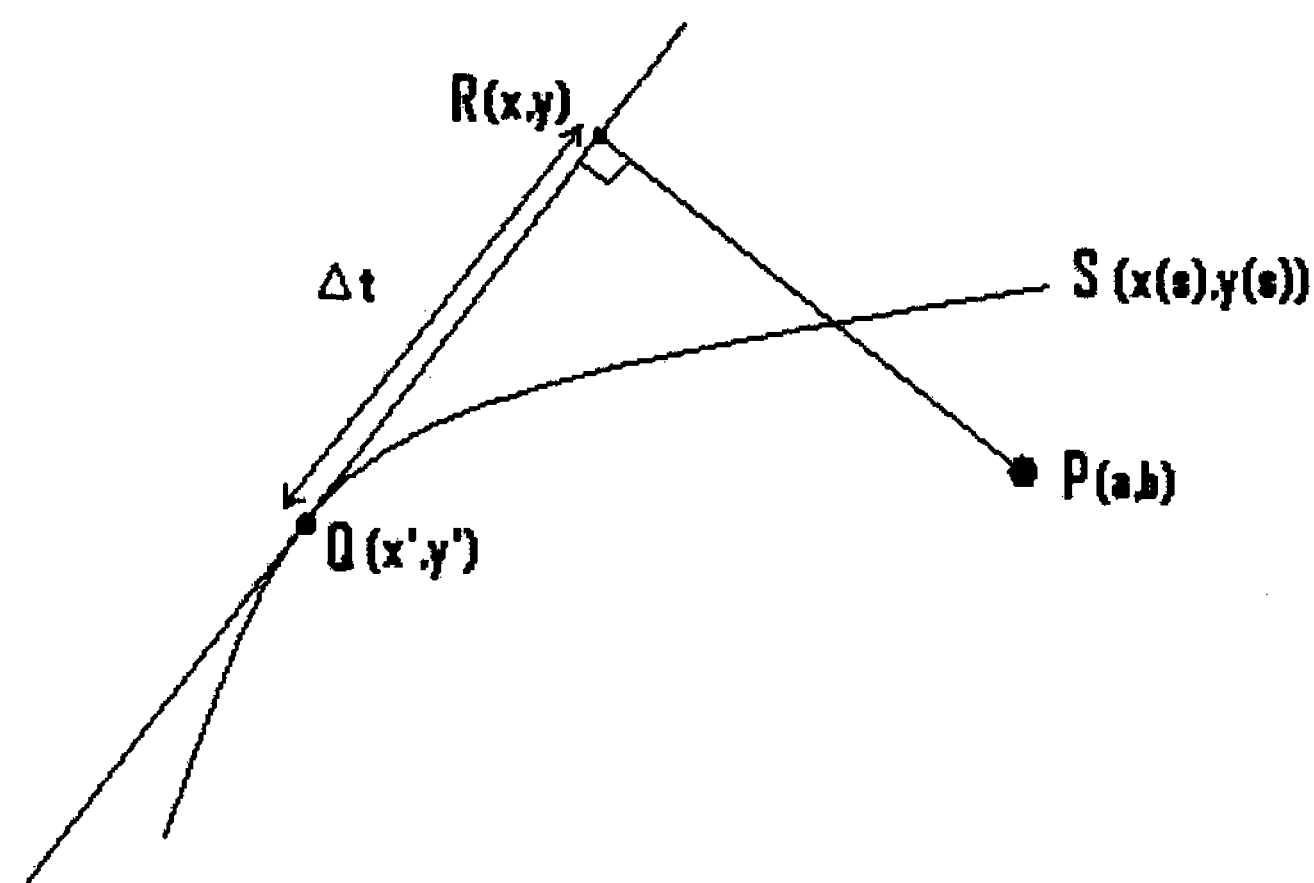


Figure 3.15: The tangent based method setup.

In the setup in Figure 3.15, 'P' is the point that needed to be projected (with known coordinates (a, b)) onto the curve 'S' which is a parameterized curve defined as below,

$$S = (x(s), y(s)) \quad \text{where } 0 \leq s \leq 1$$

Furthermore, 'Q' is any point on the parameterized curve 'S' with known coordinates (\hat{x}, \hat{y}) . The point 'R' with unknown coordinates (x, y) is placed in a such a way to make the segment 'RP' to be orthogonal to the tangent at point 'Q'. In terms of vector space, this can be stated as given in equation 3.6.

$$U_{Q-R} \cdot V_{R-P} = 0 \quad (3.6)$$

The value of Δt is the Euclidian distance between points Q and R which is given by the expression $\sqrt{(x - \hat{x})^2 + (y - \hat{y})^2}$. Furthermore, the point R can be expressed in terms of Δt , the point S_Q and its tangent (first derivative) \bar{S}_Q in the form of equation 3.7 [62],

$$R = Q + \Delta t * \bar{S}_Q \quad (3.7)$$

Present literature reveals many algorithms developed to move point Q towards minimizing Δt . Anis & Trochu attempted to converge the two points based on the equation 3.6 [63]. Here, the final resulting point is obtained by inspecting the sign (+/-) of the vector dot product, while traversing in between two end points. Yet, in our approach, a linear traversing was adopted and starting from one end (randomly selected) of the centerline, the point Q was traversed by a distance $\Delta t/2$ along the centerline. The value of $\Delta t/2$ was specifically selected based on following requirements.

1. To speed up the convergence of the point projection process.
2. To avoid overshooting the actual result by having too high a step value.

The convergence was set to be detected by setting a threshold for the Δt value. Yet, due to high shape variability of chromosomes, this method did not guarantee the global solution. Figure 3.16 depicts such a situation where the stand alone iteration method had failed. In the given scenario in Figure 3.16, the geometric iteration method was initialized from the contour point 'Q1'. Point Q2 was calculated after discovering point 'R1' on the basis of orthogonality and getting the Δt value. As

clearly visible by the projected point 'R2', the step size Δt would further reduce despite the actual point been far apart. This phenomena occurs often on chromosomes which approximately have a bend angle of 80° or above. Figure 3.17 shows such a situation on a real chromosome image in which the circle near the left end of the chromosome was the testing point for the projection and the square/star shaped dots were the iteration points (starting from the right side end of the chromosome)

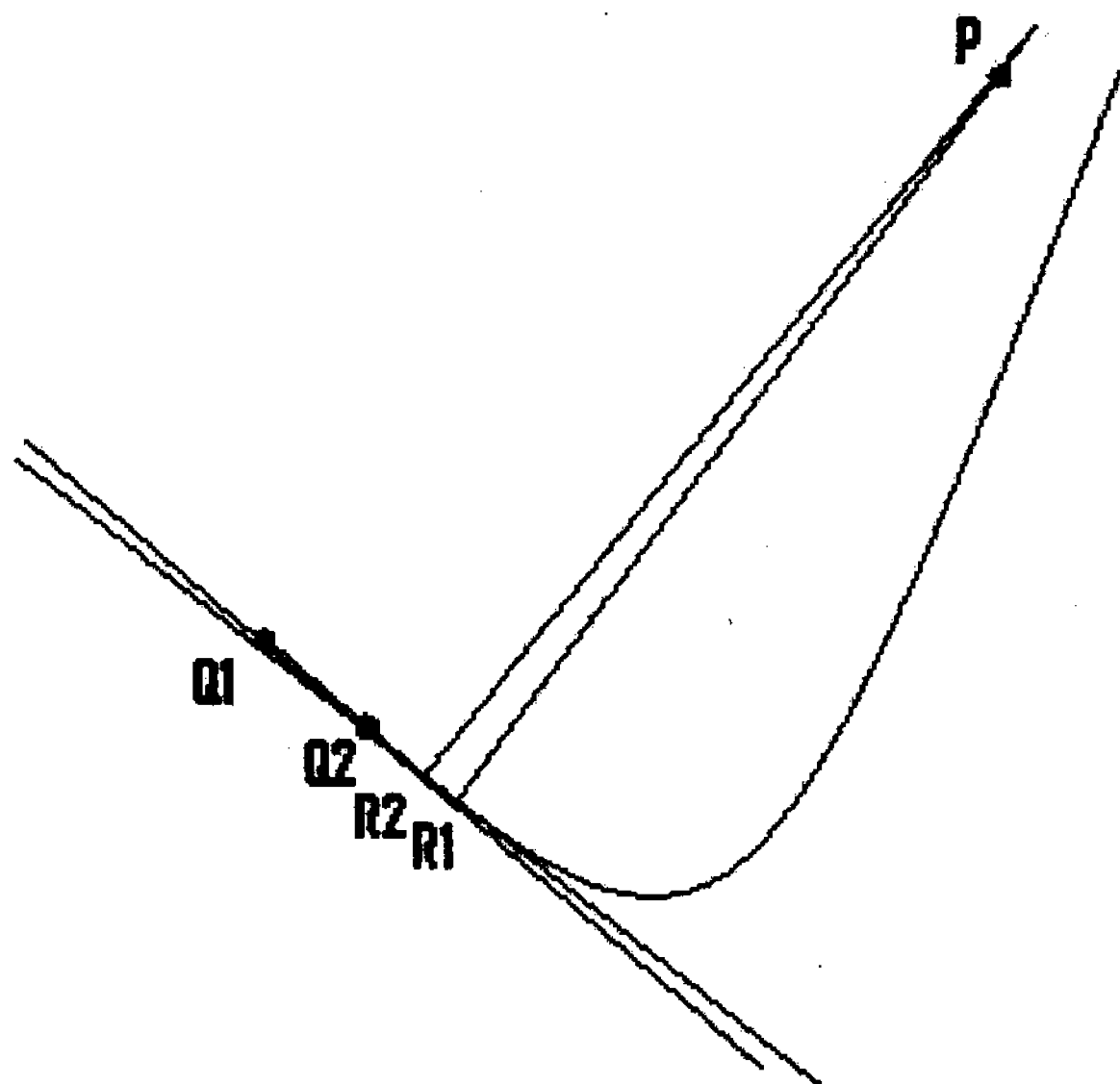


Figure 3.16: A possible fail scenario for the geometric iteration method. In this depicted situation, the point 'Q' (through Q_1, Q_2, \dots) would not converge to the point closer to the point 'P' which is the desired result.

To overcome the said limitation of the iterative method, it was required to find a suitable methodology to check the validity of the converged result. Due to the specific shape arrangement which creates this limitation, it was observed that the proposed iteration method would still perform satisfactorily provided that the algorithm was initiated at the appropriate end of the chromosome. We proposed a basic 'nearest neighbor' (NN) classification to confirm the accuracy of the iteration result. Nearest neighbor is a very intuitive classification method in which the user presumably has very little or no knowledge regarding the distribution of the test data. In our approach, the nearest neighbor method was utilized to associate one or more closest points along the centerline with the point to be projected.

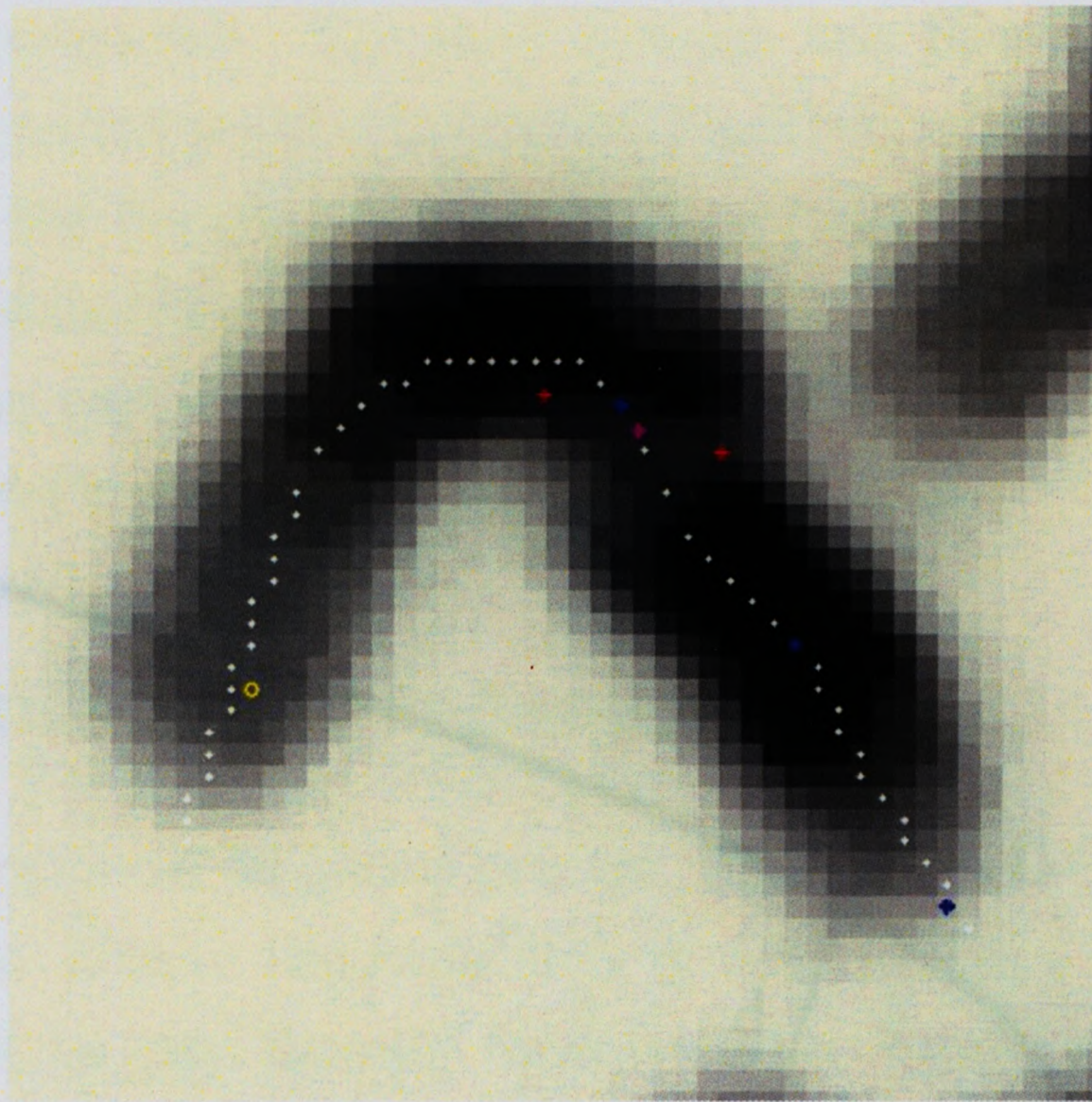


Figure 3.17: An instance where the point projection method could fail due to high bends on the chromosome, where the objective of the algorithm is to find a point closest to the 'yellow circle' on the curve. The 'blue points' shows projected points obtained at each iteration while the 'magenta' point depicting the final converged result.

The Euclidian distance was preferred over point coordinates displacement (on cartesian space). The rationale behind this preference was to make the point projection process invariable to the orientation of the chromosome. One of the main drawbacks of this method was the possibility of getting more than one point from the centerline as the closest point. Figure 3.18 depicts such a situation where three points (S_{n-1}, S_n, S_{n+1}) are placed with equal minimum displacement (d) from 'P', which is the pixel to be projected. Therefore in such situations (multiple closest points), these nearest points were organized in the proper order and the median of this series of points were obtained. This simple heuristic gave a good approximation of a benchmark (for the projected point) location which was justified experimentally on real images and test probe points.

Next, the difference (Euclidian distance) between the initial iteration method result and the nearest neighbor result was computed and this difference was thresholded (with $Thr = 4.0$). Then, if the difference of values was larger than a set threshold, the geometric iteration method was carried out starting from the other

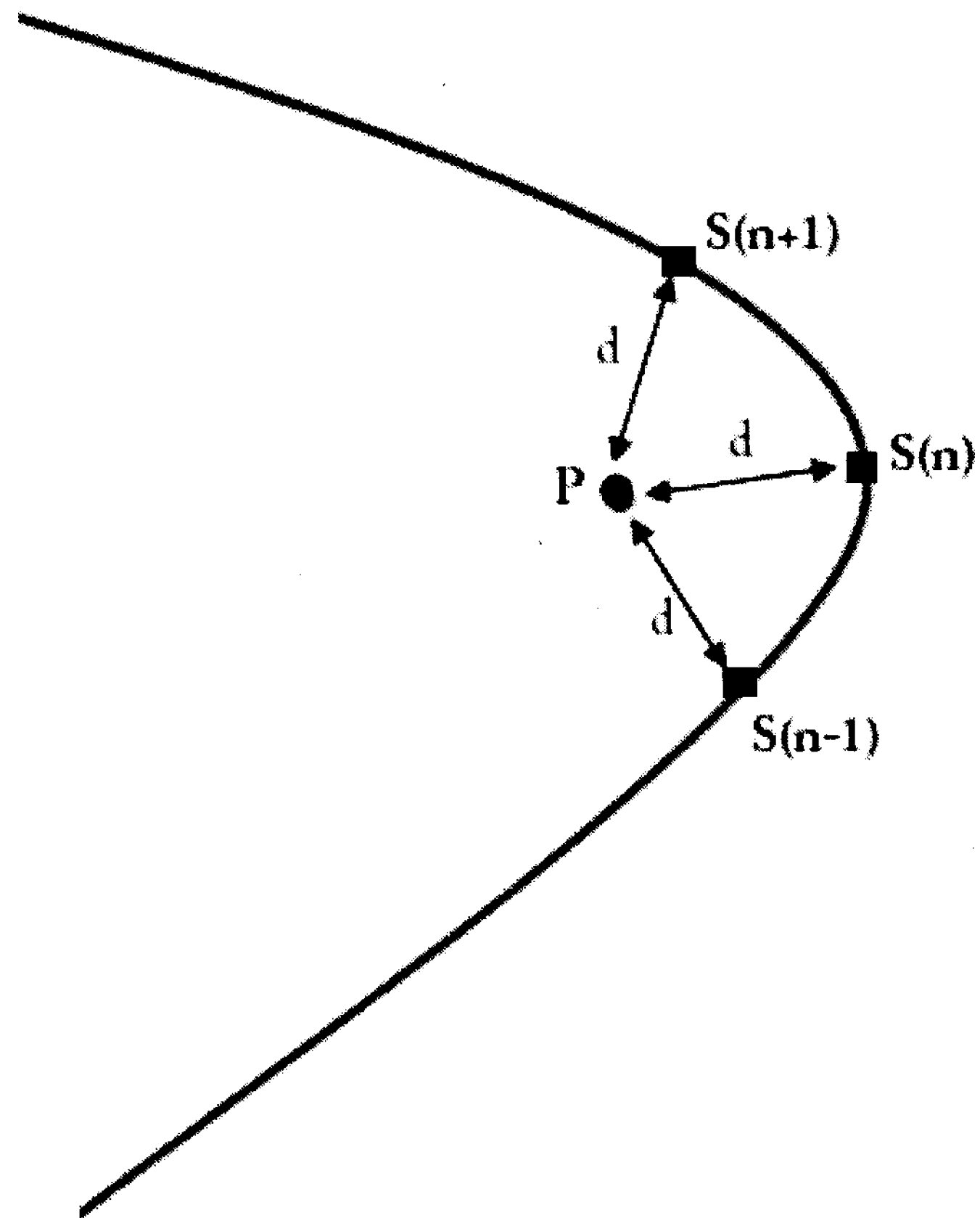


Figure 3.18: A limitation of the nearest neighbor approach where S_{n-1} , S_n & S_{n+1} are three points on the centerline with same Euclidian distance (d) from the point (FISH probe) 'P'.

end point of the chromosome. The final decision was based on both iterations in that case and the projected point was selected based on minimum difference between the NN method and the iteration results. The median point of the nearest neighbor result was not necessarily the point of interest in our case but served merely as a good approximation for the result and thus making it comparable. This point projection algorithm is illustrated in the flow chart in figure 3.19,

3.6 Fractional ratio measurement

In order to assist the clinical diagnosis process, meaningful information has to be fed to the cytogeneticists. Due to the variability in morphology of chromosomes, length measurements cannot be used directly for inferences. Thus a set of fractional ratios have to be formulated which would provide a relative value. These measures, unlike direct length measurements, can be used for clinical diagnosis.

Through the proposed algorithm, the following information/measurements are at our disposal ,

- The centerline (ends corrected) of the chromosome and therefore the Euclidian length of the chromosome (EUL)
- The centromere location of the chromosome with respect to each of the telomeres.
- The projection of the FISH probe signal onto the centerline of the chromosome. These lengths are available from both telomeres of the chromosome.

Now recall that the FISH probe only hybridizes with a known genetic sequence at a known location in the human genome. Also the centromere index (CI) value is known (with a small variation). Therefore, the relative length measurements or ratios for a 'healthy subject' are known before hand. In this research we have defined the following fractional ratio measurements M_1 , M_2 & M_3 , which can be explained using figure 3.20.

In order to define each of the above measurements, lets assume that the FISH probe falls in the Q-arm (long arm) of the chromosome as in figure 3.20. Then using figure 3.20, we can further define the following known positions,

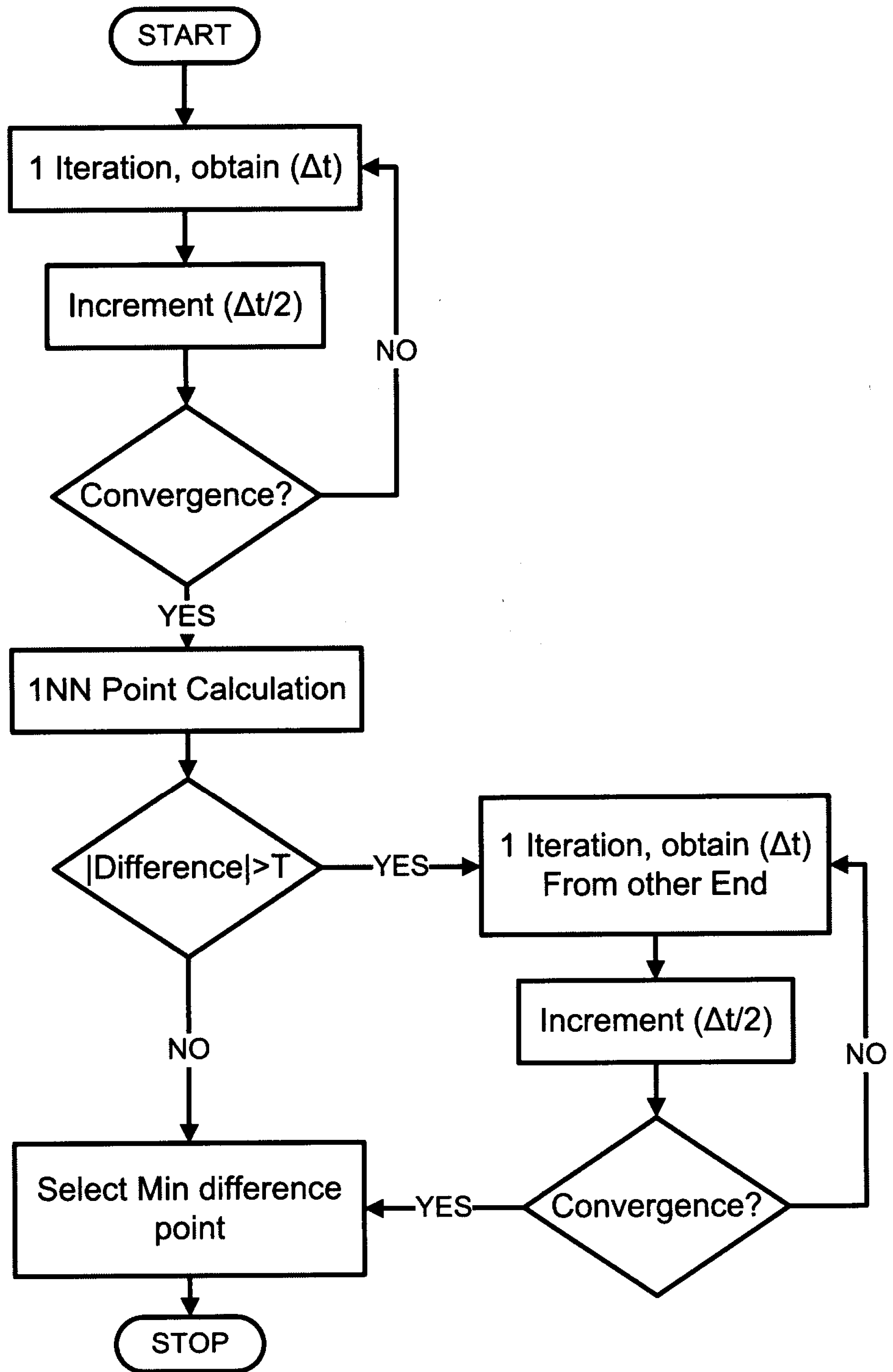


Figure 3.19: The flow chart of the point projection method used in the research.

- C_T - Centromere location of the chromosome.
- Q_T - The long arm (Q-arm) telomere of the chromosome.
- P_T - The short arm (P-arm) telomere of the chromosome.
- P - The projected point of the FISH probe signal (on the centerline).

and also the following length measurements,

- L_{QC} - The length from long arm telomere to the centromere ($Q_T \rightarrow C_T$).
- L_{PC} - The length from short arm telomere to the centromere ($P_T \rightarrow C_T$).
- L_{QP} - The length from long arm telomere to the projected point ($Q_T \rightarrow P$).
- L_{PP} - The length from short arm telomere to the projected point ($P_T \rightarrow P$).

Therefore, the following mentioned fractional ratio measurements can be calculated and defined (for each chromosome) as following.

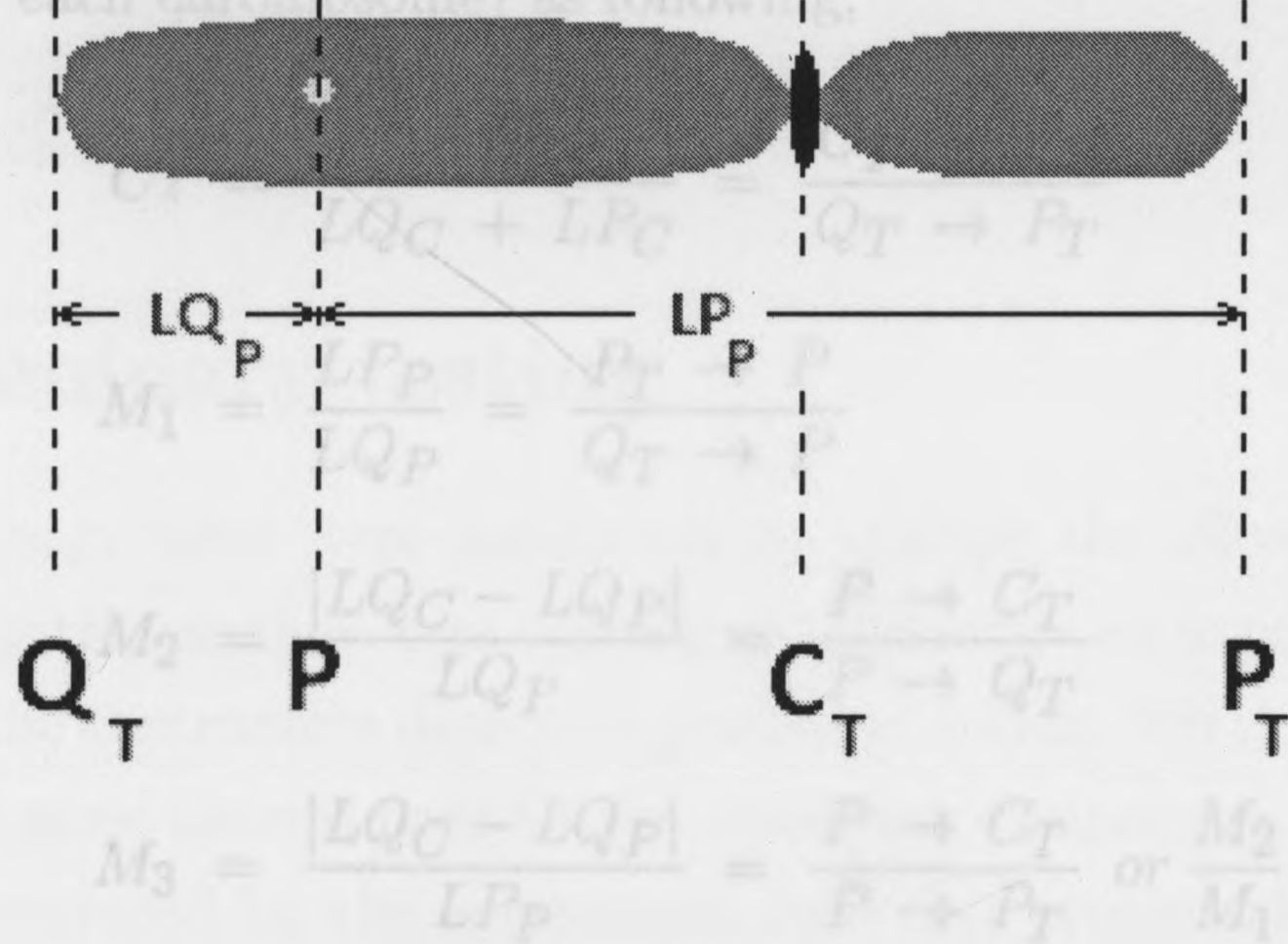


Figure 3.20: The diagram depicting the known length measurements and positions ends of the chromosome which were detected through our proposed algorithm. The value M_1 is not directly related to the centromere location. The indirect relation is from the inference of the polarity based on the centromere. Therefore, the value given by M_1 can be considered more accurate, mainly due to the difficulty in accurately detecting the centromere location.

- C_T - Centromere location of the chromosome.
 Q_T - The long arm (Q-arm) telomere of the chromosome.
 P_T - The short arm (P-arm) telomere of the chromosome.
 P - The projected point of the FISH probe signal (on the centerline).

and also the following length measurements,

- LQ_C - The length from long arm telomere to the centromere ($Q_T \rightarrow C_T$).
 LP_C - The length from short arm telomere to the centromere ($P_T \rightarrow C_T$).
 LQ_P - The length from long arm telomere to the projected point ($Q_T \rightarrow P$).
 LP_P - The length from short arm telomere to the projected point ($P_T \rightarrow P$).

Therefore, the following mentioned fractional ratio measurements can be calculated and defined (for each chromosome) as following,

$$CI = \frac{LP_C}{LQ_C + LP_C} = \frac{C_T \rightarrow P_T}{Q_T \rightarrow P_T}$$

$$M_1 = \frac{LP_P}{LQ_P} = \frac{P_T \rightarrow P}{Q_T \rightarrow P}$$

$$M_2 = \frac{|LQ_C - LQ_P|}{LQ_P} = \frac{P \rightarrow C_T}{P \rightarrow Q_T}$$

$$M_3 = \frac{|LQ_C - LQ_P|}{LP_P} = \frac{P \rightarrow C_T}{P \rightarrow P_T} \text{ or } \frac{M_2}{M_1}$$

Here, M_1 represents the fractional ratio of the probe location relative to the ends of the chromosome while M_2 & M_3 are ratios which represents the probe location with respect to the centromere location. The value M_1 is not directly related to the centromere location. The indirect relation is from the inference of the polarity based on the centromere. Therefore, the value given by M_1 can be considered more accurate, mainly due to the difficulty in accurately detecting the centromere location.

Chapter 4 Results and discussion

A novel approach for accurately detecting FISH probe locations with respect to metaphase chromosome landmarks in fluorescent microscopy images was presented in this thesis. Next, we need to quantify and analyze the improvements introduced through this algorithm. This chapter will provide the methodologies of these tests, results and a brief discussion relating to these results. The proposed algorithm was trained on 10 selected chromosomes which represented a variety of shapes and orientations. The testing of the algorithm was carried out in two stages,

1. Preliminary testing stage
2. Quantitative analysis

4.1 Preliminary testing

A set of preliminary tests were conducted to analyze the effectiveness of different stages of the algorithm as they were completed. Therefore an experiment was first carried out to test the centromere detection process in which, 306 chromosomes were extracted from the same sources used for the centerline comparison. Figure 4.2 presents typical results produced by the centromere detection process along with respective chromosome groups. Two geneticists identified the accuracy of the detected centromere location and quantified their decision into 3 different categories, namely: 'accurate', 'neighboring' and 'inaccurate'. If the detected centromere location is within 1 chromosomal band distance from the actual centromere location, the 'neighboring' label was assigned. The label 'inaccurate' was accompanied when the error is higher than 1 chromosomal band. Out of the 306 tested cases, only 10 were labeled as 'inaccurate', 257 cases were labeled as accurate and 39 cases were labeled as neighboring (refer table 4.1). These results did not depend on the chromosome type (acrocentric, sub-metacentric and metacentric) as well as on the origin of the chromosome.

Table 4.1: Expert scoring results for the centromere location detection accuracy in two data sets. Here, the term 'neighboring' is used when the detected centromere location is within 1 chromosomal band distance from the actual centromere location. The sensitivity is calculated by considering both 'accurate' and 'neighboring' as acceptable results (true positives)

Dataset	Accurate	Neighboring	Inaccurate	Cases	Sensitivity
<i>A</i>	80	1	3	84	96.4%
<i>B</i>	177	38	7	222	96.8%

Next, we tested the accuracy of the centerline extracted through our method against those of the thinning method introduced by Lam [5]. The proposed algorithm was tested on 66 chromosomes extracted from inverted DAPI stained chromosome images captured using an epifluorescence microscope. The chromosomes selected for the DCE based methodology did not overlap or touch each other and they met the minimum length criteria of 35 points (see section 3.2). The chromosomes used in this analysis came off three different individuals from six lymphocyte cells on four microscope slides. The centerline extracted using our algorithm was compared with that obtained through a thinning approach [5]. Two geneticists identified which method was better for identifying the centerline of chromosomes. Preliminary results on the expert assessment are shown in Table 4.2. In this table, chromosome images were grouped in to five classes, A, B, C, D and E by an expert. These results show that the DCE based method performs either equivalent or better than thinning in nearly all instances. Results using DCE were independent of the source of the chromosomal material, i.e. of the patient, slide or cell that was selected. The accuracy of the DCE based method was particularly high in regions of chromosome bends, which occur more frequently in longer chromosomes. Longer chromosomes are well represented in groups A, B & C in table 4.2, where the improvements of the DCE based method were apparent. In addition, longer chromosomes are found in pro metaphase chromosomes that are just beginning to condense in mitosis. Another category of relatively longer chromosomes can include chromosome rearrangements such as duplications or translocations.

The results depend on parameters of two stages: namely, the GVF segmentation and the sampling point selection. Firstly, the segmentation outcome was observed to be highly sensitive to the values set for the main internal parameters of the GVF snake such as α (elasticity factor), β (rigidity factor), μ (GVF regularization factor)

Table 4.2: Comparative scoring of DCE vs thinning algorithms - where each value gives the number of chromosomes for which, the centerline was better represented by the corresponding algorithm

Class	DCE	Thinning	Both	Total
<i>A</i>	7	-	5	12
<i>B</i>	5	-	5	10
<i>C</i>	19	-	9	28
<i>D</i>	3	1	4	08
<i>E</i>	6	-	2	08

and κ (external force weight). A sensible set of values for the above factors ($\alpha = 0.05$, $\beta = 0$, $\kappa = 2$, $\mu = 0.2$) provided satisfactory results in our experiments but fine tuning was possible through adjustment. Next, the selection of the sample point spacing and the starting and ending sample point offset (from the DCE result) were observed as critical parameters in the centerline extraction and values were set based on empirical observations.

The centerline through our algorithm was also observed to be able to successfully handle bent chromosomes (as seen in Fig 4.1) and extracted centerline that closely represented the shape information of the chromosome. The sharpness of the centerline results in Fig 4.1 is mainly due to the pixelation effect of the magnified image.

A preliminary analysis of probe detection by point projection was also carried out. A test probe position was arbitrarily assigned on a chromosome to observe and compute the projected point on the centerline. The initial testing of the point projection on chromosome images yielded satisfactory results upon visual inspection (see figure 4.3). A more qualitative test has to be performed in the future to further verify these observations.

The confidence measure circumscribing the predicted centromere will be used during probe localization to weight the contribution of this feature during chromosome abnormality detection. Probe localization on chromosomes with low confidence centromere placement could be biased towards the relative distances to the termini of the chromosome. Alternatively, a model-based approach for centromere analysis can be combined with these confidence intervals. This would involve testing the chromosome-specific probe for detecting a sequence in its normal context, where the expected location of the centromere is already known.

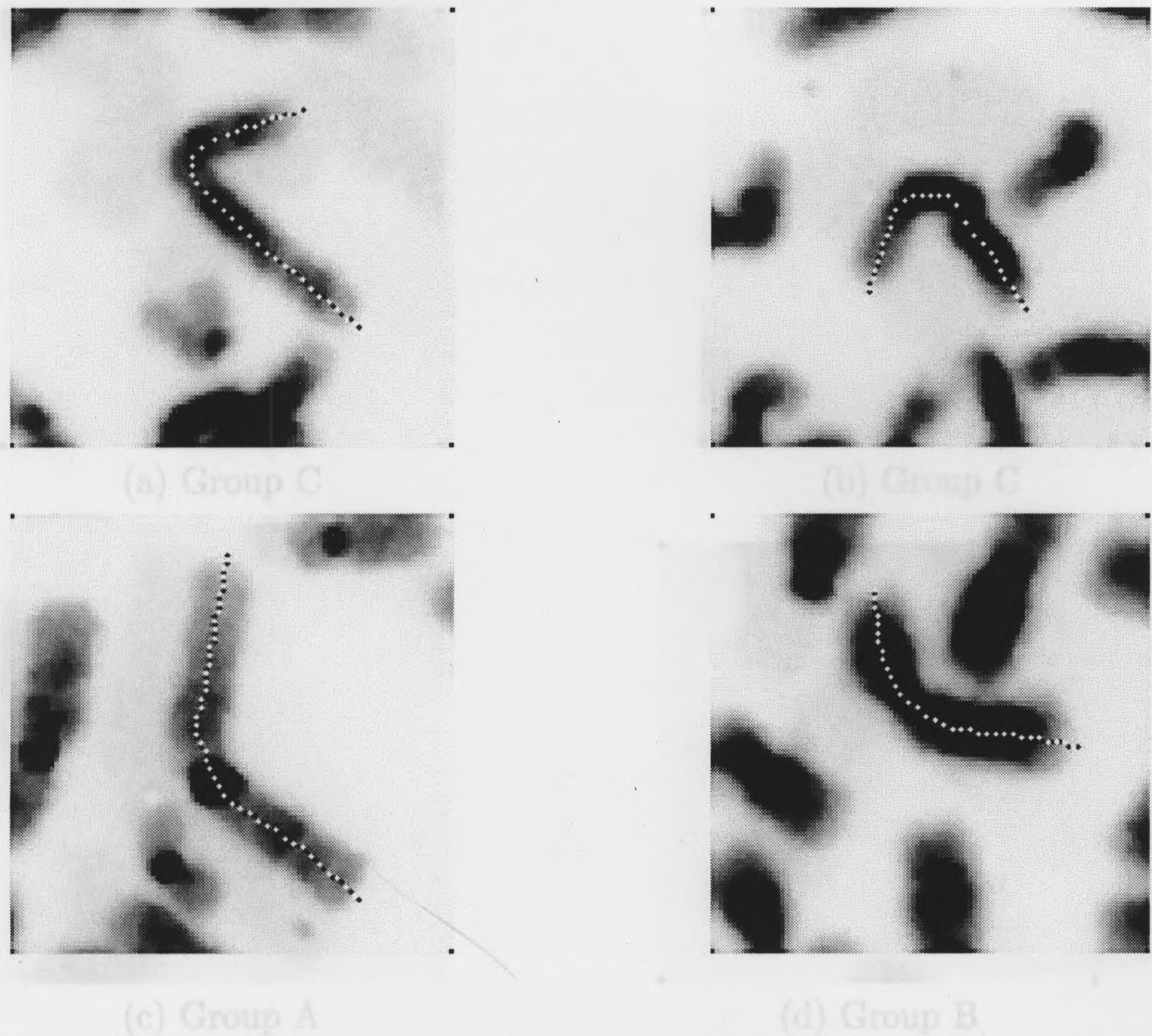


Figure 4.1: End point corrected centerline (black & white line) results of representative bent chromosomes using the proposed approach.

4.2 Quantitative analysis

A proper quantitative analysis was needed to establish a platform to analyze and compare accuracy of the proposed algorithm. This was achieved by fitting a metric into one of the prominent features obtained through the algorithm. It was noted earlier that the accuracy of the extracted centerline as a landmark, heavily determines the outcome of any other measurement result on the chromosome. Therefore, the centerline of the chromosome was selected as the feature to fit into the error metric. The thinning method described by Lam [5] was used for comparing the results of the proposed method. A centerline manually recorded by an expert was used as the 'gold standard' in the analysis.

Yet, certain practical issues arose when analyzing the accuracy of a centerline with comparison to the 'gold standard'.

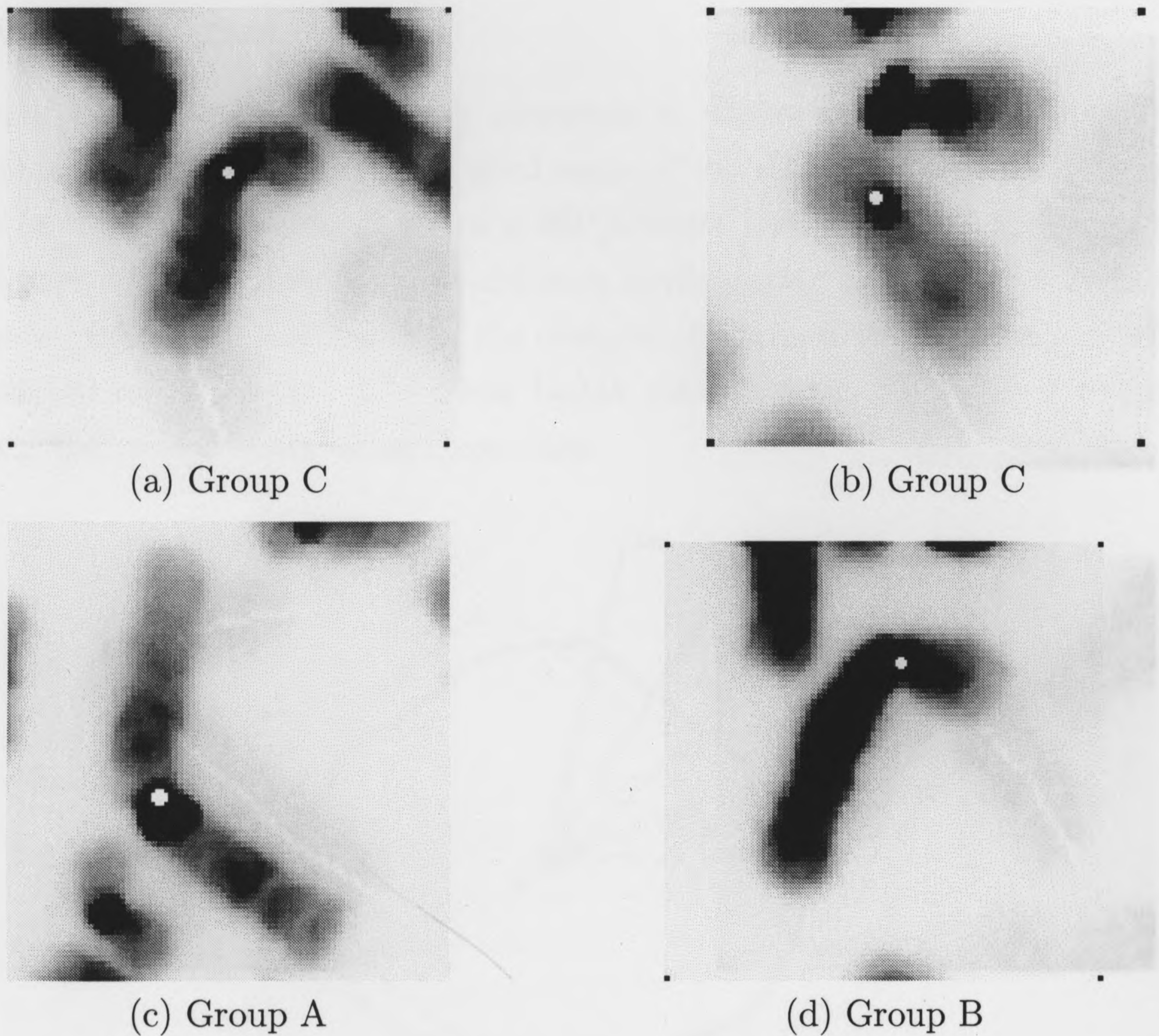


Figure 4.2: Some of the chromosome centromere detection results with their respective chromosome groups.

- A centerline drawn by an expert can still be subjective. This is due to the 'inter-' and 'intra-observer' variability introduced by human operators. One remedy is to provide the same set of data to many experts in random order as performed by some experiments [65].
- It could also be difficult to decide a metric that would properly reflect the accuracy of the algorithm.
- The use of different metrics and subjective gold standards can make different methods incomparable.
- Collecting large sets of data and gold standards are tedious and time consuming.

Therefore, the 'gold standard' used in our analysis can be biased and user-

4.2.1 Error metric

Selecting a proper reference point is a key operation in analyzing the performance of any algorithm. In this paper we have used some of the distance error metrics proposed by Ladak in analyzing the segmentation of a 2D prostate boundary [66]. Ladak used a configuration of line segments starting from the centroid (C), creating an angle θ with the selected reference point $C - C'$. Then Ladak measured the distance $d(\theta)$ (length) between the two contours along those lines.

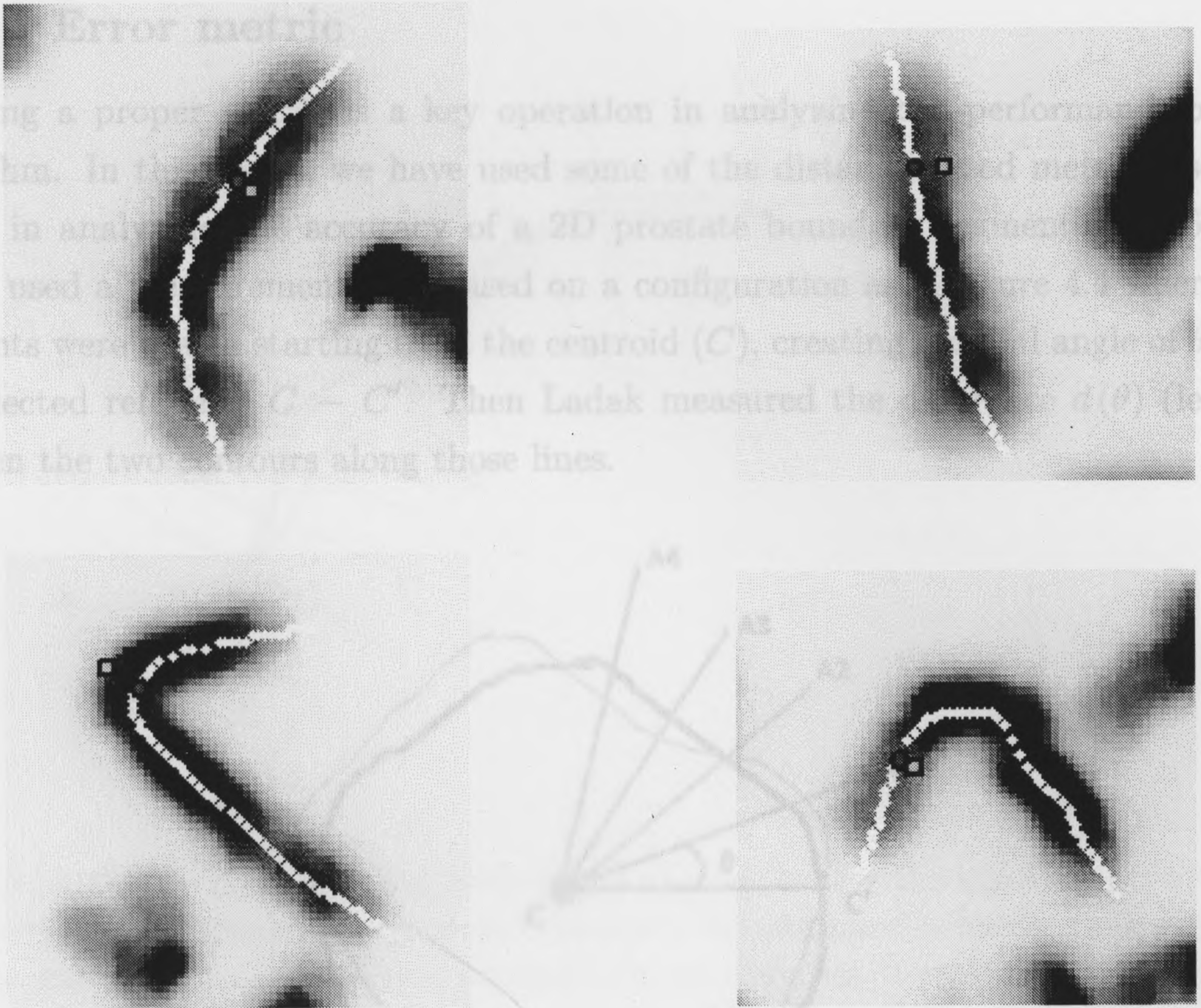


Figure 4.3: Point projection results - (yellow square - test point : blue circle - projected point)

dependent. But we could compensate this to a certain extent by using the centerline obtained through thinning as a benchmark.

The gold standard contour and the red contour represents the gold standard and the actual segmentation respectively [6]. The measurement value $d(\theta)$ was carried out along the line segments ($C - A1$ etc.) based on the radial angle value θ (between $C - C'$ and $C - A1$)

Our interest was in analyzing the difference between two curve segments. Therefore, the centroid held no meaning as a reference point. Therefore in our analysis, each gold standard point was taken iteratively as a reference point. Then, for each of these points, a perpendicular line was created based on the local gradient. The distance error measure $d(i)$ was then measured along these line segments (or in the proximity of 2 pixels) as depicted in figure 4.5.

Let the gold standard curve segment be denoted by $G(i) = (x_i, y_i)$, where

4.2.1 Error metric

Selecting a proper metric is a key operation in analyzing the performance of any algorithm. In this thesis, we have used some of the distance based metrics used by Ladak in analyzing the accuracy of a 2D prostate boundary segmentation [6],[66]. Ladak used a measurement grid based on a configuration as in figure 4.4 where line segments were drawn starting from the centroid (C), creating a radial angle of θ with the selected reference $C - C'$. Then Ladak measured the difference $d(\theta)$ (length) between the two contours along those lines.

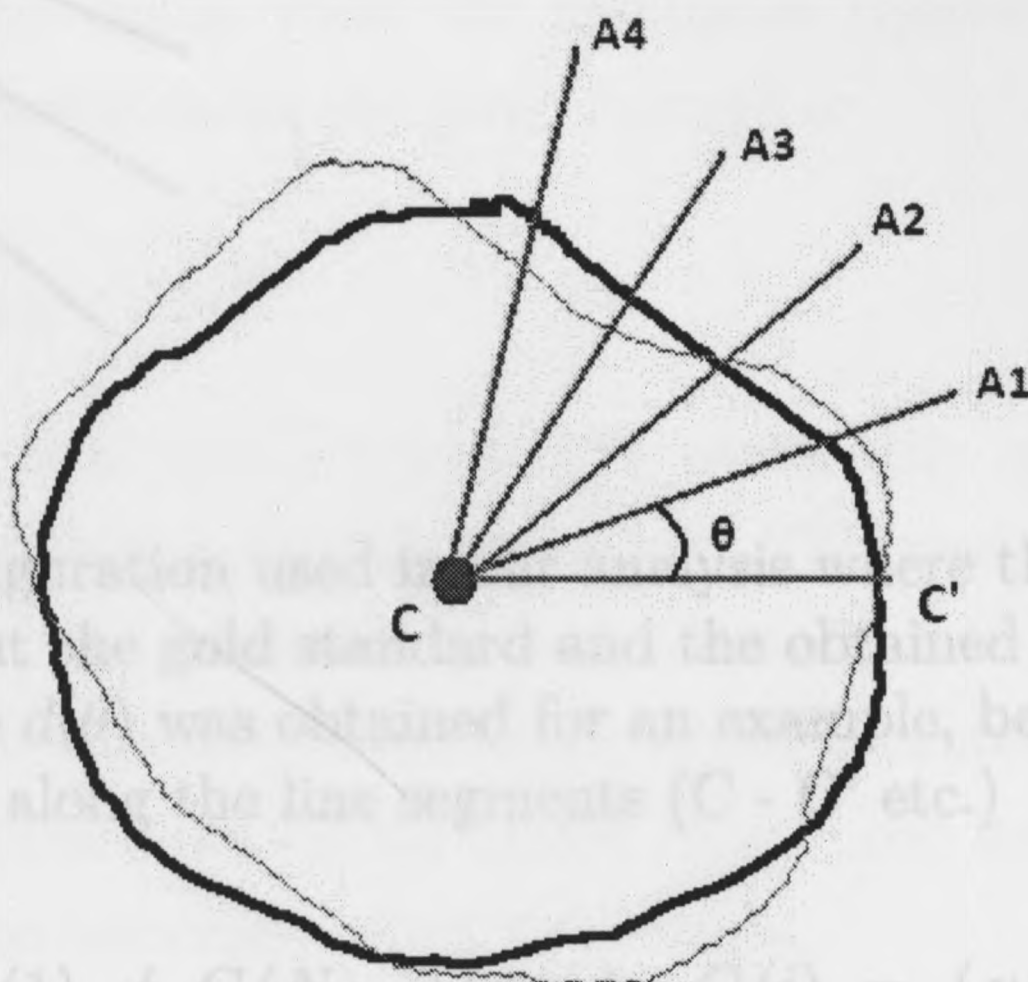


Figure 4.5: The configuration used by Ladak to analyze accuracy of a 2D prostate boundary segmentation where the black and the red curve segments represent the gold standard and the obtained centerline respectively. The measurement value $d(\theta)$ was obtained for an example, between points P and R, along the line segments (C - etc.)

$i = \{1, 2, \dots, N\}$ and $G(1) \neq G(N)$. Let $C(i) = (x_i, y_i)$ be the computer generated contour in which, each index i corresponds to that of the gold standard. Then the distance $d(i)$ for each point can be simply calculated as shown below.

Figure 4.4: The configuration used by Ladak to analyze accuracy of a 2D prostate boundary segmentation where the black contour and the red contour represents the gold standard and the actual segmentation respectively [6]. The measurement value $d(\theta)$ was carried out along the line segments (C - A1 etc.) based on the radial angle value θ (between C-C' and C-A1)

Our interest was in analyzing the difference between two curve segments. Therefore, the centroid held no meaning as a reference point. Therefore in our analysis, each gold standard point was taken iteratively as a reference point. Then, for each of these points, a perpendicular line was created based on the local gradient. The distance error measure $d(i)$ was then measured along these line segments (or in the proximity of 2 pixels) as depicted in figure 4.5.

Let the gold standard curve segment be denoted by $G(i) = (x_i, y_i)$, where

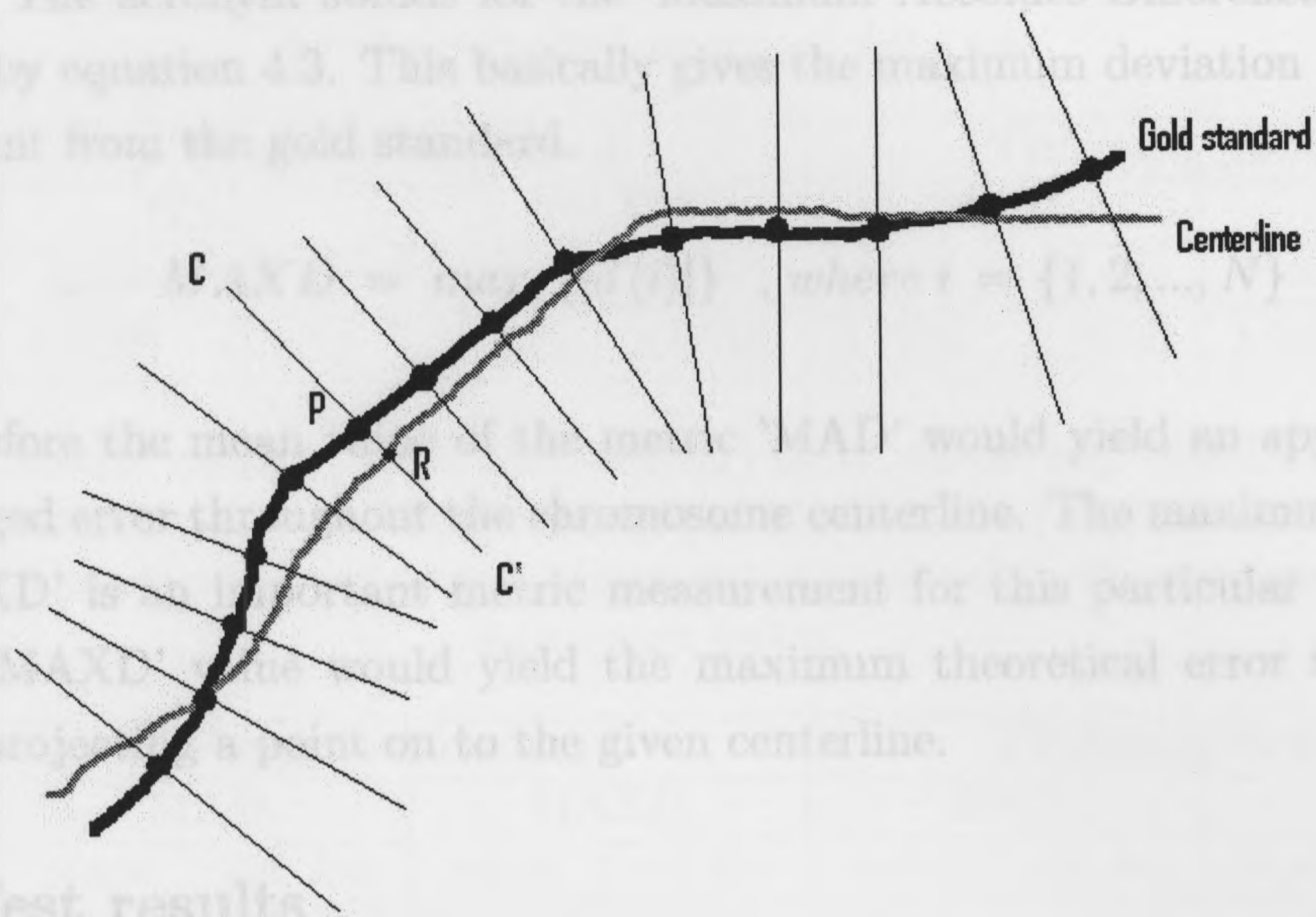


Figure 4.5: The configuration used in our analysis where the black and the red curve segments represent the gold standard and the obtained centerline respectively. The measurement value $d(\theta)$ was obtained for an example, between points P and R, along the line segments (C - C' etc.)

Also, due to inter and intra-observer variability introduced into the gold standard by a single expert could also adversely affect the results. $i = [1, 2, \dots, N]$ and $G(1) \neq G(N)$. Then, let $C(i) = (u_i, v_i)$ be the computer generated contour in which, each index i corresponds to that of the gold standard. Then the distance $d(i)$ for each point can be simply calculated as shown below,

$$d(i) = \sqrt{(x_i - u_i)^2 + (y_i - v_i)^2} \quad (4.1)$$

After getting the value of $d(i)$ for each point, the two error metrics can be defined as follows,

MAD : This stands for 'Mean Absolute Distance' and is defined by equation 4.2. This metric takes the absolute difference into account and then average it. Therefore, the value of the MAD yields the mean value of the error, disregarding the sign of the error.

$$MAD = \frac{1}{N} \sum_{i=1}^N |d(i)| \quad (4.2)$$

MAXD : The acronym stands for the 'Maximum Absolute Difference' and is defined by equation 4.3. This basically gives the maximum deviation of the curve segment from the gold standard.

$$MAXD = \max_i \{|d(i)|\} , \text{ where } i = \{1, 2, \dots, N\} \quad (4.3)$$

Therefore the mean value of the metric 'MAD' would yield an approximation of an averaged error throughout the chromosome centerline. The maximum deviation value 'MAXD' is an important metric measurement for this particular application. Here, the 'MAXD' value would yield the maximum theoretical error that can be caused by projecting a point on to the given centerline.

4.2.2 Test results

It is important to note that this setup doesn't necessarily reflect the effects or the presence of spurious branches in the thinned result. This was one of the major factors considered for scoring at the preliminary stage centerline testing setup discussed in section 4.1. Therefore, the metric values of the thinned centerline can be unrealistic. Also, due to inter and intra-observer variability introduced into the gold standard by a single expert could also adversely affect the results.

A total of 120 metaphase chromosomes taken from 15 lymphocyte cells were used for this stage of testing. The selection of chromosomes were carried out to meet the following criteria,

- Chromosomes that do not overlap or touch any other chromosome in the cell image.
- Each of the extracted chromosomes were required to satisfy the minimum length criteria of 35 skeletal points used for the DCE based approach in our hybrid solution. If this criteria was not met, then our algorithm would select a morphological thinning algorithm that would yield results identical to the thinning result used for comparison.

Table 4.3 provides the mean and standard deviation of metric values (MAD and MAXD) for both methods based on 120 metaphase chromosomes. The results

Table 4.3: Matric results (MAD and MAXD) mean values and standard deviation for the complete analyzed data set.

No. of Chromosomes	Proposed method				Thinning method			
	MAD		MAXD		MAD		MAXD	
	<i>mean</i>	<i>std dev</i>	<i>mean</i>	<i>std dev</i>	<i>mean</i>	<i>std dev</i>	<i>mean</i>	<i>std dev</i>
120	0.6300	0.2725	1.7643	0.9536	0.7278	0.3220	2.0071	1.1557

Table 4.4: Matric results (MAD and MAXD) mean values and standard deviation for each chromosome group. The number (within brackets) after each chromosome group name specifies the number of chromosomes present in the data set from that particular group.

Chromosome Group	Proposed method				Thinning method			
	MAD		MAXD		MAD		MAXD	
	<i>mean</i>	<i>std dev</i>	<i>mean</i>	<i>std dev</i>	<i>mean</i>	<i>std dev</i>	<i>mean</i>	<i>std dev</i>
A (21)	0.5952	0.2965	1.6849	0.8866	0.6735	0.3127	1.8961	1.1025
B (22)	0.6613	0.2227	1.6849	0.5242	0.7164	0.3001	1.9849	0.8945
C (55)	0.6217	0.2405	1.8612	1.1664	0.7299	0.2794	2.1153	1.2613
D (11)	0.6811	0.2644	1.5207	0.5061	0.7902	0.3686	1.8572	0.9517
E (09)	0.7315	0.4849	1.7285	0.6236	0.8936	0.5597	1.9507	1.4897
F (01)	0.3016	-	1.4142	-	0.4905	-	1.4142	-
G (01)	0.6243	-	1.4142	-	0.5828	-	1.4142	-

in table 4.3 implies a decrease in both error metrics MAD and MAXD values for the proposed method when compared to those with the thinning method.

Next, the table 4.4 depicts the error metric values for chromosome groups A to G. From figure 4.7 and table 4.4, it can be observed that the mean values of both metrics (MAD and MAXD) are less with our proposed method than with the thinning method, for all the chromosome groups from 'A' to 'E'. Conclusions cannot be drawn on the error rates of groups 'F' and 'G' due to the limited number of analyzed samples from those groups. Through figure 4.6 and table 4.4, it can also be seen that the margin of variation (standard deviation) of our proposed method yields better results than the benchmark thinning method. The approximately constant mean error metric values also suggests a uniform performance of the proposed algorithm across different chromosome groups (refer figure 4.7). Yet, more samples are needed to verify these claims. Another possible verification could be obtained by testing against another expert 'gold standard' to rule out inter-observer variability.

4.2.3 Variabilities or dependencies

The morphology of a DAPI stained chromosome could vary due to numerous variabilities and it is important to test our proposed algorithm against some of them. By doing so, the robustness of the algorithm can be explored.

4.2.3.1 Based on the patient

In this stage, chromosomes obtained from three different patients were analyzed. Table 4.5 consists the resulting error metric values (mean and std. deviation) of the two metrics for these three patients. The mean and std. deviation of the metric 'MAD' depicts uniform performance within the tested data set. Yet, the values of 'MAXD' standard deviation seems to be inconsistent for the three patients. It is important to observe that this inconsistency is also evident on the thinning method results. Therefore, it is not possible (with the analyzed data set) to arrive at a conclusion for this dependance.

Table 4.5: Matric results (MAD and MAXD) mean values and standard deviation for 16 chromosomes from patient No.1, 21 chromosomes from patient No.2 and 10 chromosomes from patient No.3.

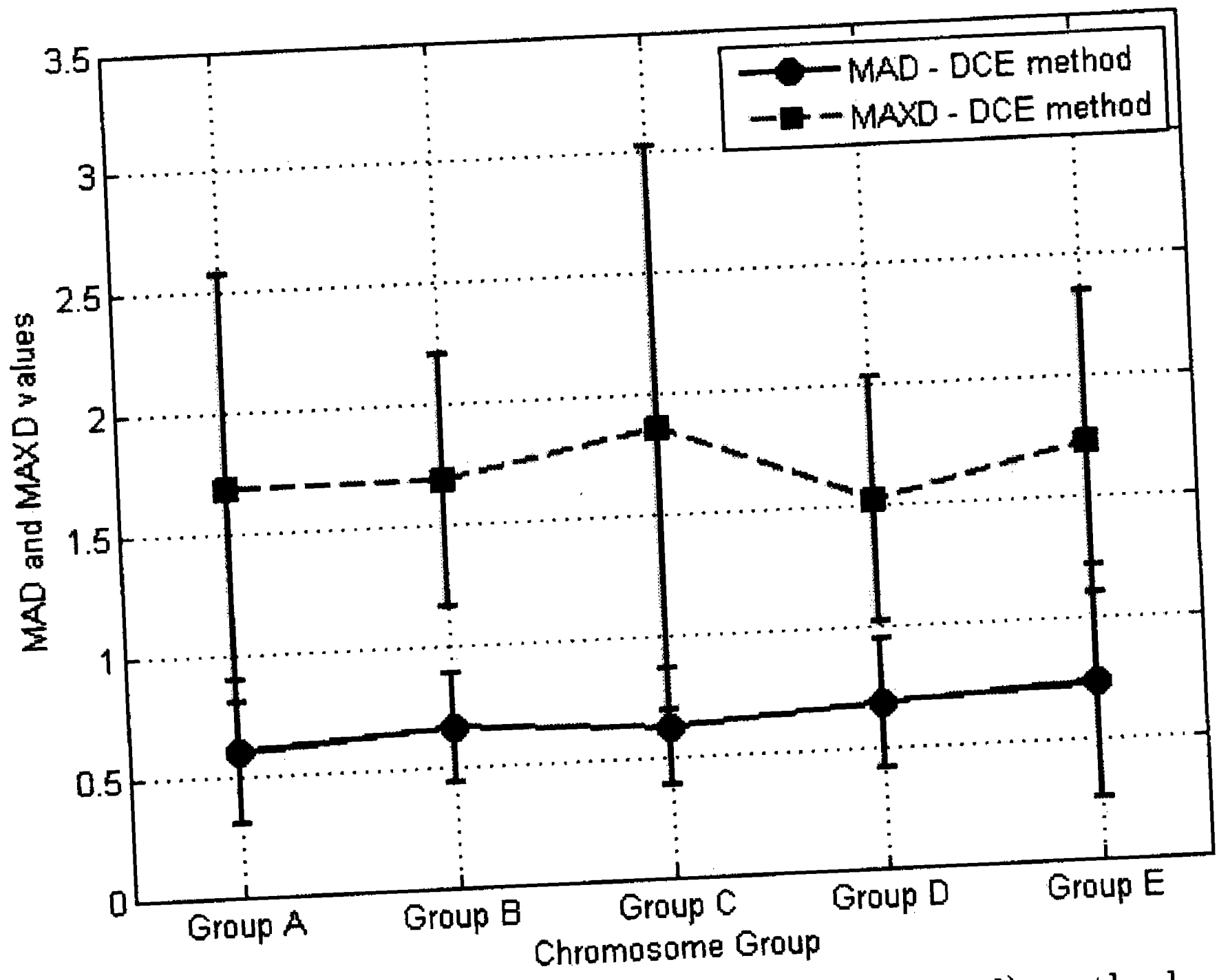
Patient number	Proposed method				Thinning method			
	MAD		MAXD		MAD		MAXD	
	<i>mean</i>	<i>std dev</i>	<i>mean</i>	<i>std dev</i>	<i>mean</i>	<i>std dev</i>	<i>mean</i>	<i>std dev</i>
No.1	0.5851	0.2335	1.9508	1.4401	0.6340	0.3527	2.2196	1.5953
No.2	0.5482	0.2361	1.6965	0.4773	0.6505	0.2434	1.8838	0.8798
No.3	0.5317	0.1973	1.3728	0.1310	0.5487	0.2351	1.4721	0.3842

4.2.3.2 Based on the slide

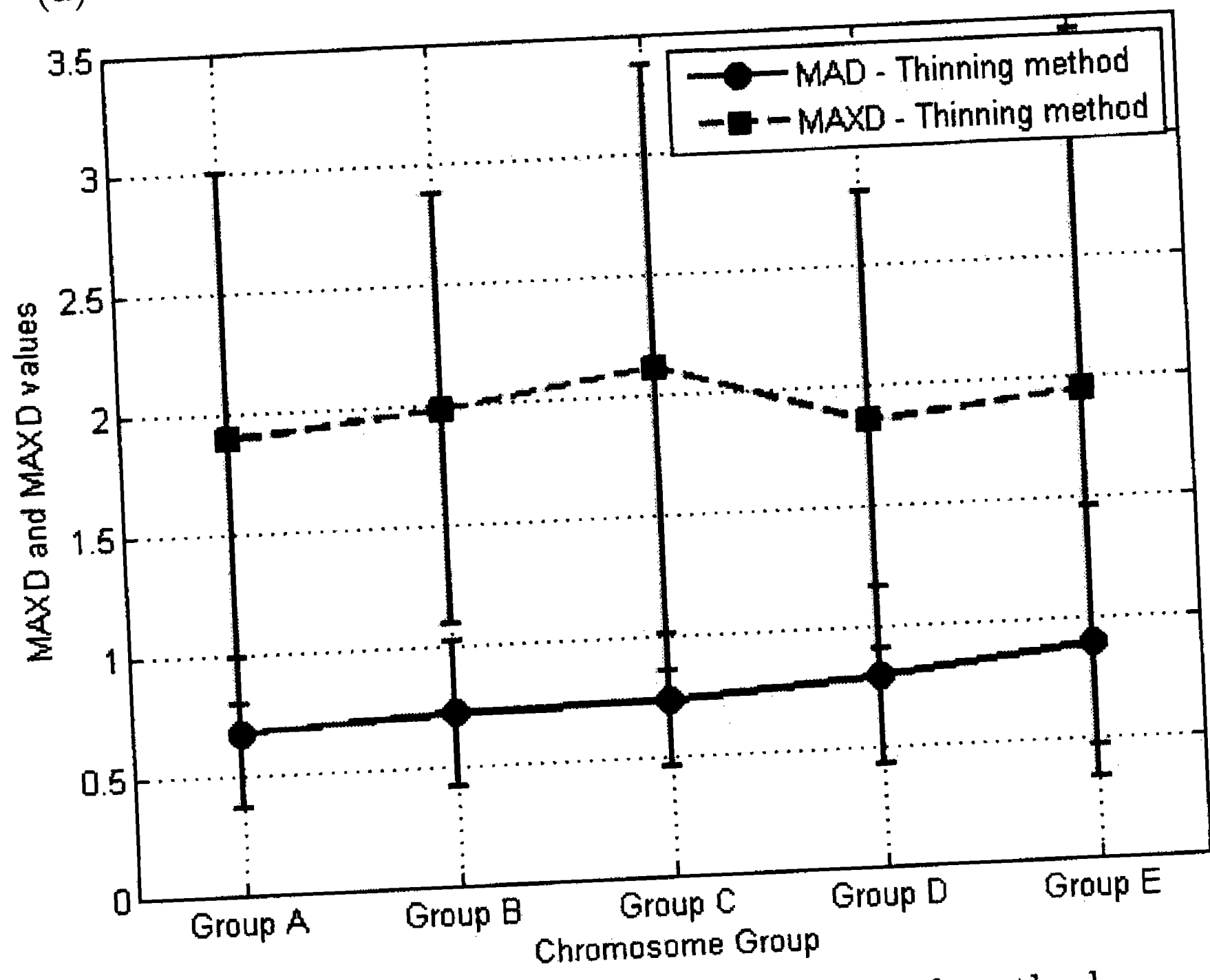
Here, the performance on metaphase chromosomes from different slides of a single patient is analyzed and the results are depicted in table 4.6. It can be observed that the results for each metric value of the proposed algorithm is consistent for both slides of the same patient. A further analysis on a larger data set has to be performed to verify this observation.

Table 4.6: Matric results (MAD and MAXD) mean values and standard deviation for 10 chromosomes from slide No.1 and 11 chromosomes from slide No.2 from the same patient.

Slide number	Proposed method				Thinning method			
	MAD		MAXD		MAD		MAXD	
	<i>mean</i>	<i>std dev</i>	<i>mean</i>	<i>std dev</i>	<i>mean</i>	<i>std dev</i>	<i>mean</i>	<i>std dev</i>
No.1	0.5472	0.2477	1.8485	0.4509	0.6707	0.3086	1.9715	0.9506
No.2	0.5492	0.2372	1.5584	0.4781	0.6321	0.1789	1.8040	0.8485

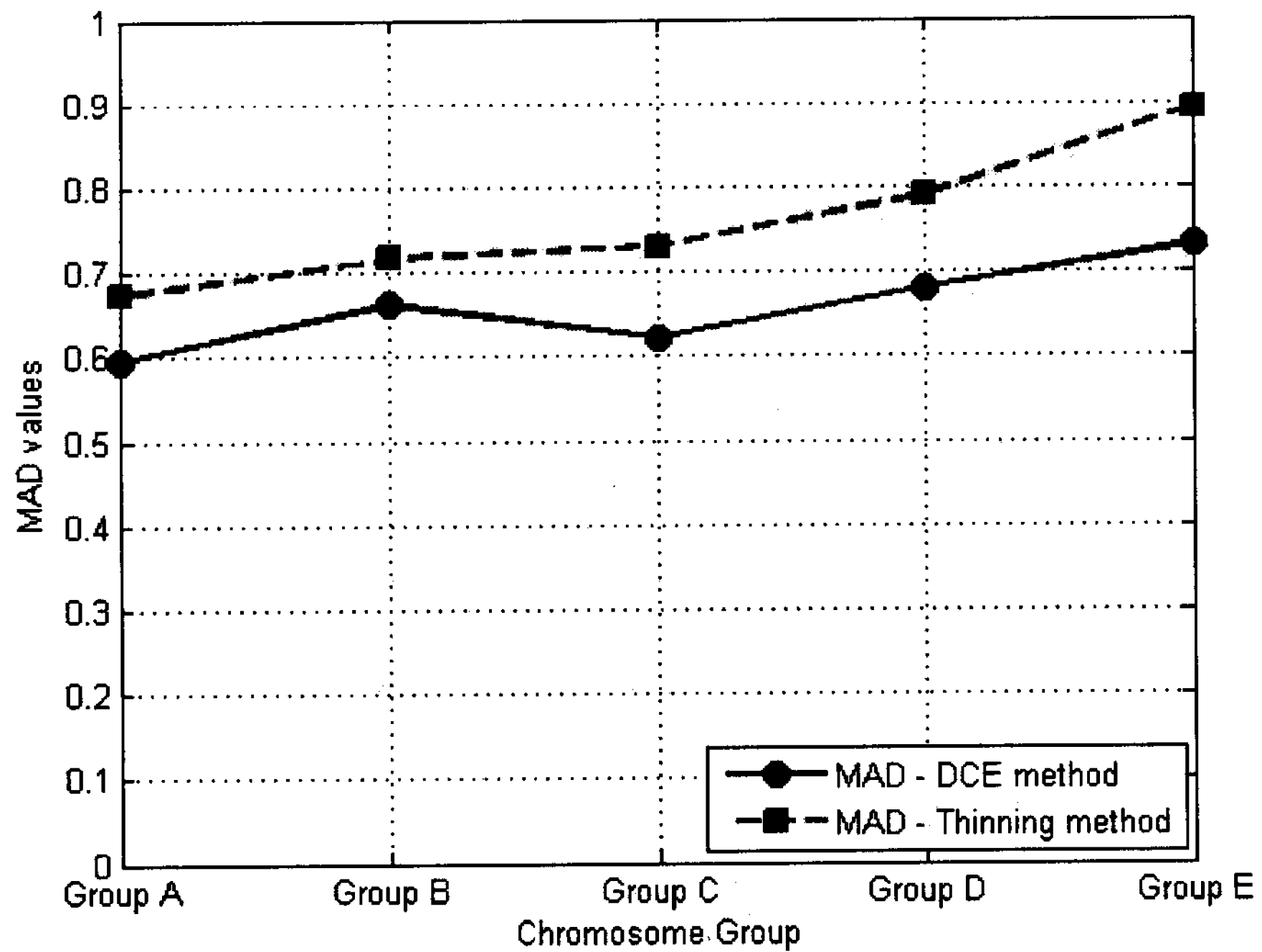


(a). Metric results of the 'DCE' based (proposed) method

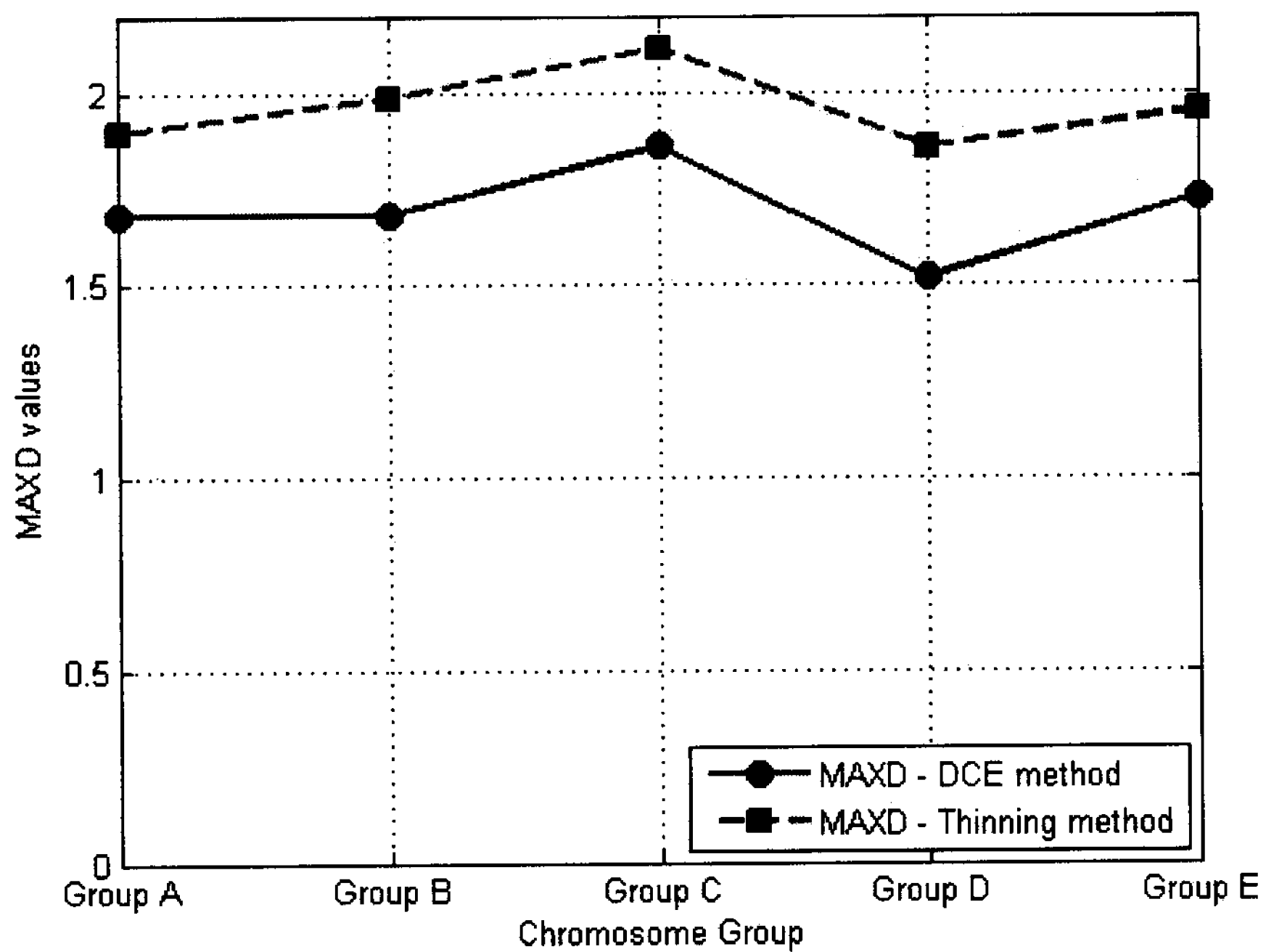


(b). Metric results of the 'thinning' based method

Figure 4.6: MAD and MAXD (centerline) metric mean and std. deviation values of chromosome groups A to E.



(a). MAD metric results for the proposed (DCE) and thinning methods



(b). MAXD metric results for the proposed (DCE) and thinning methods

Figure 4.7: Comparison of MAD and MAXD (centerline) metric mean values of chromosome groups A to E obtained through the proposed (DCE based) and thinning methods.

Chapter 5 Conclusion and future work

An algorithm that utilizes GVF active contours, DCE based skeleton pruning and morphological thinning which locates probe signal in relation to chromosome landmarks is presented in this thesis. This algorithm has been tested on real metaphase chromosome images as discussed in chapter 4. The proposed method yielded a 13.43% (mean of the MAD metric) accuracy improvement (calculated from table 4.3) in comparison to the morphological thinning method [5]. This chapter will provide some conclusive remarks and also will discuss some of the feasible future work.

5.1 Summary and conclusion

A semi-automatic hybrid algorithm was proposed to make fractional ratio measurements based on the FISH probe signals on metaphase chromosome images. The proposed algorithm mainly combines gradient vector flow snakes, discrete curve evolution based skeleton pruning, morphological thinning and cubic spline interpolation. The ability of this algorithm to accurately detect main features of chromosomes such as the centerline and the centromere were tested in chapter 4. The current state of the proposed algorithm (given the coordinate of the FISH probe signal) can automatically perform the following operations,

- A proper segmentation of the chromosome of interest. i.e. - the chromosome with the FISH probe signal.
- Obtaining a smooth skeleton and then pruning spurious branches.
- Gradient and intensity based telomere detection of the chromosome.
- Detecting the centromere location of the chromosome based on the previously calculated centerline. Then incorporating the confidence of this detection in a novel measure termed as the 'centromere confidence' measure (CCF value).
- Projecting the FISH probe signal onto the centerline and then making 3 fractional measurements based on it (see section 3.6).

The proposed algorithm is robust against image boundary noise as well as the high variability of the chromosome shapes. The ability to project FISH probe signals to an accurate centerline approximation, is an important stage in developing a computer based setup to assist clinical diagnosis. Our algorithm can be readily adopted for FISH probe signal localization on chromosome images. This algorithm can be extended effectively to applications outside the domain of chromosome images. One such application in geography could be finding the line that connects the highest altitude points on a group of mountains.

5.2 Future work

Future work includes a suitable method for splitting overlapping chromosomes along with more improvements for the end point correction method to make this algorithm applicable to any metaphase chromosome. Also some refinement can be made to the introduced 'centromere confidence measure' (CCF) value. One such method would be to consider a range of values instead of the absolute width profile minimum.

Furthermore, a methodology needs to be developed to accurately extract a small FISH probe signal (with a known spectral frequency) from a noisy digital image with lots of information. The signal strength of the probe signal itself, is of varying strength depending on the exposure time setting of the camera. A background subtraction method based on Gabor filtering is currently being explored in order to obtain an accurate representation of the probe signal and by doing so, to make the process fully autonomous.

This algorithm can be tested on a larger data set to draw more conclusions on its accuracy. Furthermore, some testing has to be performed to analyze the point projection accuracy of this algorithm. These results can then be compared with the expected ratios of a healthy person. Further analysis of such a large set of data and there probe ratio result could prove to be productive.

References

- [1] M. Y. Karşlıgil, M. Elif & Karşlıgil, *Fuzzy Similarity Relations for Chromosome Classification and Identification*, ser. Lecture Notes in Computer Science. Springer Berlin / Heidelberg, January 1999, vol. 1689, pp. 142 – 148, .CAIP 99.
- [2] C. Xu and J. L. Prince, “<http://iacl.ece.jhu.edu/projects/gvf>,” a web site for the GVF snake demos and examples.
- [3] L. J. Latecki and R. Lakämper, “Convexity rule for shape decomposition based on discrete contour evolution,” *Computer Vision and Image Understanding*, vol. 73, no. 3, pp. 441 – 454, March 1999.
- [4] C. Hilditch, “Linear skeletons from square cupboards,” in *Machine Intelligence*, vol. 4. Edinburgh Univ. Press, 1969, pp. 403 – 420.
- [5] L. Lam and S. W. Lee, “Thinning methodologies-a comprehensive survey,” *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, vol. 14, no. 09, pp. 869 – 885, September 1992.
- [6] N. Hu, D. Downey, A. Fenster, and H. Ladak, “Prostate boundary segmentation from 3d ultrasound images.” *Med Phys*, vol. 30, no. 7, pp. 1648–59, 2003.
- [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. Natl. Acad. Sci. USA*, vol. 85, pp. 9138–9142, December 1988.
- [8] P. Rogan and J. Knoll, “<http://www.cytognomix.org/public.html>,” website for Laboratories of Genome Bioinformatics and Genomic Disorders.
- [9] J. Piper and E. Granum, “On fully automatic feature measurement for banded chromosome classification,” *Cytometry*, vol. 10, pp. 242–255, 1989.
- [10] 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.

- [11] G. Ritter and G. Schreib, "Using dominant points and variants for profile extraction from chromosomes," *Pattern Recognition Journal*, no. 4, pp. 923–938, April 2001.
- [12] J. Graham et al., "Automatic karyotype analysis," *Chromosome Analysis Protocols*, vol. 29, pp. 141–185, 1994.
- [13] M. Popescu et al., "Automatic karyotyping of metaphase cells with overlapping chromosomes," *Computers in Biology and Medicine*, vol. 29, no. 1, pp. 61–82(22), January 1999.
- [14] S. L. Gerson and M. B. Keagle, *The Principles of Clinical Cytogenetics*, 2nd ed. Humana Press, 2005.
- [15] R. King et al, *A dictionary of genetics*, 5th ed. Oxford university press, 1968.
- [16] T. Kobayashi et al., "Content and classification based ranking algorithm for metaphase chromosome images," in *IEEE Conference on Multimedia Imaging*, 2004.
- [17] W. Qiang et al., *Microscope Image Processing*. ELSEVIER Academic Press, 2008.
- [18] M. Moradi et al., "Automatic locating the centromere on human chromosome pictures," in *16th IEEE Symposium on Computer-Based Medical Systems*, 2003.
- [19] 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.
- [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] V. Gajendran and J. Rodriguez, "Chromosome counting via digital image analysis," in *International Conference on Image Processing(ICIP)*, October 2004, pp. 24–27.

-
- [22] J. Canny, "A computational approach to edge detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, vol. 8, no. 6, November 1986.
- [23] L. Ji, "Fully automatic chromosome segmentation," *Cytometry*, vol. 17, pp. 196–208, 1994.
- [24] 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.
- [25] X. Wang et al., "Automated classification of metaphase chromosomes: Optimization of an adaptive computerized scheme," *Journal of Biomedical Informatics*, vol. 42, no. 01, pp. 22 – 31, February 2009.
- [26] G. Enrico et al., "Automatic segmentation of chromosomes in q-band images," in *Proceedings of the 29th Annual International Conference of the IEEE EMBS*, August 2007, pp. 23–26.
- [27] M. Kass et al., "Snakes: Active contour models," *International Journal of Computer Vision*, vol. 1, no. 4, pp. 321–331, January 1988.
- [28] C. Xu and J. L. Prince, "Gradient vector flow: A new external force for snakes," in *Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 1997.
- [29] 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.
- [30] 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*, 2005.
- [31] 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, no. 06, March 1990.

- [32] 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.
- [33] S. Biswas and B. C. Lovell, *Chapter 9: Snakes and Active Contours - Bzier and Splines in Image Processing and Machine Vision*. Springer London.
- [34] C. Xu and J. L. Prince, "Snakes, shapes, and gradient vector flow," *IEEE Transaction on Image Processing*, vol. 7, no. 3, 1998.
- [35] 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.
- [36] C. Xu and J. L. Prince, *HANDBOOK OF MEDICAL IMAGING: Processing and Analysis*, I. Bankman, Ed. Academic Press, 2000.
- [37] L. D. Cohen, "On active contour models and balloons," in *CVGIP: Image Understanding archive*, vol. 53. Academic Press, Inc, 1991, pp. 211 – 218.
- [38] 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.
- [39] L. J. Latecki and R. Lakämper, "Contour-based shape similarity," in *Visual Information and Information Systems*, ser. Lecture Notes in Computer Science, vol. 1614/1999. Springer Berlin, January 1999, p. 657.
- [40] L. Latecki and R. Lakämper, "<http://knight.cis.temple.edu/shape/shape/index.html>," web site.
- [41] 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.
- [42] X. Bai et al., "Skeleton pruning by contour partitioning with discrete curve evolution," *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, vol. 29, no. 03, March 2007.

- [43] 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*. Springer-Verlag London, UK, 1999, pp. 398 – 409.
- [44] L. J. Latecki et al., "Continuity of discrete curve evolution," *Journal of Electronic Imaging*, vol. 09, no. 03, July 2000.
- [45] K. Siddiqi and A. Shokoufandeh, "<http://www.cim.mcgill.ca/shape/>," web site.
- [46] B. Grewal, *Higer Engineering Mathematics*, sixteenth ed. Khanna Publishers, Delhi-110 006, India, July 2001.
- [47] C. Runge, "Über empirische funktionen und die interpolation zwischen äquidistanten ordinaten," *Zeitschrift für Mathematik und Physik*, vol. 46, pp. 224 – 243, 1901.
- [48] J. W. Ferguson, "Multivariable curve interpolation," *Journal of the ACM (JACM)*, vol. 11, no. 02, pp. 221 – 228, 1964.
- [49] G. Farin, *Curves and Surfaces for Computer Aided Geometric Design : A Practical Guide*, 3rd ed. Academic Press, INC., 1993.
- [50] D. Salomon, *Curves and Surfaces for Computer Graphics*. Berlin, Germany / Heidelberg, Germany: Springer-Verlag, 2006.
- [51] J. White, "Micromaths cubic spline curve fitting," *Teaching Mathematics and Its Applications*, vol. 05, no. 01, pp. 39 – 45, 1986, .MEP Curriculum Development Project, West Sussex Institute of Higher Education.
- [52] D. F. Rogers and J. A. Adams, *MATHEMATICAL ELEMENTS for COMPUTER GRAPHICS*, 2nd ed. New York, NY, USA: McGraw-Hill, Inc., 1989.
- [53] G. H. Behforooz, "A comparison of the e(3) and not-a-knot cubic splines," *Applied Mathematics and Computation*, vol. 72, no. 2-3, pp. 219 – 223, October 1995.
- [54] C. De Boor, "Convergence of cubic spline interpolation with not-a-knot condition," *Mathematic Research Center, University of Wiscosin*, October 1985, pre-print.

- [55] M. Kozubek, *Image Acquisition and its Automation in Fluorescence Microscopy*. Springer Netherlands, 2006, vol. 3.
- [56] Mathworks, “www.mathworks.com/access/helpdesk/help/toolbox/images/,” web site.
- [57] Gonzalez and Woods, *DIGITAL IMAGE PROCESSING*, 2nd ed. Prentice Hall.
- [58] H. Blum, “A transformation for extracting new descriptors of shape,” in *Models for the Perception of Speech and Visual Form*. MIT Press, 1967, pp. 362 – 380.
- [59] D. Shaked and A. Bruckstein, “Pruning medial axes,” *Computer Vision and Image Understanding*, vol. 69, no. IV970598, pp. 156–169, 1998.
- [60] 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, no. 04, April 2002.
- [61] E. R. Faria et al., *Segmentation and Centromere Locating Methods Applied to Fish Chromosomes Images*. Springer Berlin / Heidelberg, 2005.
- [62] S. M. Hu and J. Wallner, “A second order algorithm for orthogonal projection into curves and surfaces,” in *Computer Aided Geometric Design*, vol. 22. Elsevier Science Publishers, March 2005, pp. 251 – 260.
- [63] A. Limaiem and F. Trochu, “Geometric algorithms for the intersection of curves and surfaces,” in *Computer & Graphics*, vol. 19, no. 03. Elsevier Science Publisher, March 1995, pp. 391 – 403.
- [64] E. Hartmann, “On the curvature of curves and surfaces defined by normalforms,” *Computer Aided Geometric Design*, vol. 16, pp. 355 – 376, March 1999.
- [65] H. M. Ladak, Y. Wang, D. B. Downey, and A. Fenster, “Testing and optimization of a semiautomatic prostate boundary segmentation algorithm using virtual operators,” *Medical Physics*, vol. 30, no. 7, pp. 1637 – 1647, 2003.
- [66] A. C. Hodge, A. Fenster, D. B. Downey, and H. M. Ladak, “Prostate boundary segmentation from ultrasound images using 2d active shape models: Optimisation and extension to 3d,” *Comput. Methods Prog. Biomed.*, vol. 84, no. 2-3, pp. 99 – 113, 2006.