The Dissertation Committee for Wade Carl Schwartzkopf certifies that this is the approved version of the following dissertation:

Maximum Likelihood Techniques for Joint Segmentation-Classification of Multi-spectral Chromosome Images

Committee:

Brian L. Evans, Supervisor

Alan C. Bovik, Co-Supervisor

Kenneth R. Castleman

Joydeep Ghosh

Thomas E. Milner

John A. Pearce

Maximum Likelihood Techniques for Joint Segmentation-Classification of Multi-spectral Chromosome Images

by

Wade Carl Schwartzkopf, B.S.C.S.E.E., M.S.E.E.

Dissertation

Presented to the Faculty of the Graduate School of

The University of Texas at Austin

in Partial Fulfillment

of the Requirements

for the Degree of

Doctor of Philosophy

The University of Texas at Austin

December, 2002

This dissertation is dedicated to my grandfather, Carl Corder.

Maximum Likelihood Techniques for Joint Segmentation-Classification of Multi-spectral Chromosome Images

Publication No._____

Wade Carl Schwartzkopf, Ph. D.

The University of Texas at Austin, 2002

Supervisors: Brian L. Evans and Alan C. Bovik

This dissertation develops new methods for automatic chromosome identification by taking advantage of the multispectral information in M-FISH chromosome images and by jointly performing chromosome segmentation and classification. Chromosome imaging is a valuable tool for doctors and cytogenetic technicians. Extra chromosomes, missing chromosomes, broken chromosomes, and translocations (parts of chromosomes breaking off and attaching to other chromosomes) are indicators of radiation damage, cancer, and a wide variety of inherited diseases. There are currently over 325 clinical cytogenetics laboratories in the United States performing over 250,000 diagnostic studies each year involving chromosome analysis.

Traditional chromosome imaging has been limited to grayscale images, but recently a 5-fluorophore combinatorial labeling technique (M-FISH) was

developed in which each class of chromosomes binds with a different combination of fluorophores. This results in a multi-spectral image, in which each class of chromosomes has distinct spectral components. Although M-FISH presents significantly more information than was available in traditional grayscale images, little research on multispectral chromosome image analysis has been previously reported in the open literature.

The purpose of the research described in this dissertation is to develop new methods for automatic chromosome identification. In particular, I (1) develop a maximum likelihood hypothesis test that uses this multi-spectral information, together with conventional criteria, to select the best segmentation possibility, (2) use this likelihood function to combine chromosome segmentation and classification into a robust chromosome identification system, and (3) show that the proposed likelihood function can also be used as a reliable indicator of errors in segmentation, errors in classification, and the chromosomes anomalies that can be diagnosed with M-FISH imaging. I show that the proposed multi-spectral joint segmentation-classification method outperforms past grayscale segmentation methods in decomposing touching chromosomes. Furthermore, I show that it outperforms past M-FISH classification techniques that do not use segmentation information.

Table of Contents

List of Tables

List of Figures

Chapter 1: Introduction

Chromosomes are the structures in cells that contain genetic information. When chromosomes are photographed during cell division, the images of these chromosomes contain much information about the health of an individual. Chromosome images are useful for diagnosing genetic disorders and for studying various diseases, such as cancer. In the past it was necessary for laboratory technicians to examine these images visually. This manual process of locating, classifying, and evaluating the chromosomes in these images could be lengthy and tedious. Since visual inspection is time consuming and expensive, and since many images often have to be inspected, many attempts have been made to automate the analysis of these chromosome images. Many algorithms have been developed to assist the laboratory technician in locating and classifying chromosomes. Computers are now indispensable tools in cytogenetic laboratories, but full, automated image analysis is still quite far away. However, the practical goal is to reduce the user interaction time of the laboratory technician either by segmenting and classifying more chromosomes accurately or by aiding the user in manual segmentation and classification.

In the mid-1990's, a new technique for staining chromosomes was introduced. It produced an image in which each chromosome type appeared to be a distinct color [1]. This multi-spectral staining technique made analysis of chromosome images easier, not only for visual inspection of the images by humans, but also for computer analysis of the images. The multispectral staining

technique is called M-FISH (multiplex fluorescence in-situ hybridization.) M-FISH uses five color dyes that attach to various chromosomes differently to produce a multi-spectral image, and a sixth dye that attaches to all chromosomes to produce a grayscale image. This dissertation develops a method to take advantage of the color information in M-FISH images to improve on past methods of computer analysis of chromosome images. It introduces a probabilistic model of M-FISH chromosomes which can be used for simultaneous segmentation and classification.

1.1 OBJECTIVES

This work investigated what improvements in chromosome analysis could be obtained by using M-FISH multi-spectral images. One aim of this work was to develop algorithms that could take full advantage of the multi-spectral information and to quantify their improvement over grayscale chromosome image analysis methods. In order to achieve effective segmentation and classification, I studied how to combine segmentation and classification to make both more accurate. Furthermore, this work examined how to use this multi-spectral information to detect automatically abnormalities in the chromosomes that could not be detected without such chromosome labeling.

1.2 CONTRIBUTIONS

The best approach for segmentation and classification of multi-spectral chromosome images is fundamentally different from the best approach for grayscale chromosome images. The additional information provided by multi-spectral chromosome images is useful for distinguishing touching and

overlapping chromosomes within clusters and thus for segmenting them from one another. Therefore the best approach for segmentation of multi-spectral chromosome images is not simply to apply traditional techniques to a grayscale version of the image. *This dissertation defends the idea that joint segmentation-classification based on optimization of probabilistic information obtained from the multi-spectral chromosome pixels enables decomposition of touching and overlapping chromosomes and provides estimates of the confidence in the chromosome segmentation-classification.*

Specifically, the contributions of this dissertation are as follows:

1. A maximum likelihood hypothesis test is proposed as a method for selecting the best way to decompose groups of chromosomes that touch and overlap each other. An algorithm is described which efficiently uses this criterion in the multi-dimensional color-space that M-FISH images use. Finally, results of this algorithm are summarized and compared with that of past methods of chromosome image analysis.
2. This maximum likelihood test is used to propose a method that combines the task of locating and classifying chromosomes for improved performance in both tasks.
3. The first two contributions in the dissertation are then used to achieve aberration scoring; that is, giving a score to each segment to indicate the likelihood of abnormalities in that image.

Table 1.1: Acronyms used in this dissertation

ADIR	Advanced Digital Imaging Research
ASI	Applied Spectral Imaging
DAPI	4',6-Diamidino-2-phenylindole
DNA	deoxyribonucleic acid
ISCN	International System for Human Cytogenetic Nomenclature
M-FISH	multiplex fluorescence in-situ hybridization
ML	maximum likelihood

1.3 NOTATION

This section gives the mathematical notation that is followed throughout the dissertation. I denote vectors with boldface and scalars with plain font. $\mathbf{x}(\mathbf{m}) \in \Re^n$ refers to a multi-spectral n dimensional vector pixel at two-dimensional location $\mathbf{m} \in Z^2$. For convenience, sometimes I drop the explicit two-dimensional indices and denote a vector pixel as $\mathbf{x}$.

I use the function $p(\cdot)$ to denote a probability. Further, the function $G(\cdot,\cdot,\cdot)$ denotes the usual Gaussian probability density function. Thus the probability density function of a random vector $\mathbf{x}$, with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$, is $G(\mathbf{x}, \boldsymbol{\mu}, \boldsymbol{\Sigma})$.

Table 1.1 lists the acronyms used in this dissertation.

1.4 ORGANIZATION

The structure of this dissertation is as follows. Chapter 2 introduces the features of chromosomes and chromosomes images and shows how these features

can be used to classify chromosomes into a karyotype. The basics of the M-FISH multi-spectral chromsome imaging technique are also described. The prior research in chromosome image analysis is reviewed and the potential application of image analysis techniques to M-FISH images is examined.

In Chapter 3, I describe my formulation of the M-FISH segmentation-classification problem and introduce a maximum likelihood hypothesis test for evaluating the quality of a given segmentation. The likelihood proposed is a function of both segmentation and classification, and thus can be used to evaluate both simultaneously. I then propose maximizing this function to perform segmentation and classification simultaneously.

Chapter 4 describes a practical method for effectively selecting a set of reasonable segmentation possibilities for evaluation by the proposed hypothesis test. It begins by attempting to oversegment the image and then combines pairs of segments in a way that most increases the proposed likelihood function. Pairs continue to be combined until no more pairs are found that increase the likelihood function.

I present results of my algorithm in Chapter 5. I show that the proposed multi-spectral joint segmentation-classification method outperforms past grayscale segmentation methods in decomposing touching chromosomes, and I show that it outperforms past M-FISH classification techniques that do not use segmentation information. In addition, I show that the proposed likelihood function is a reliable indicator of abnormal chromosomes, as well as segmentation and classification errors.

Finally, Chapter 6 reviews the contributions of the dissertation and suggests directions for future research.

Chapter 2: Background

2.1 INTRODUCTION

This chapter introduces the basic terminology and concepts used in this dissertation. It introduces chromosome imaging, features useful for classifying chromosomes in images, and some physical disorders that can be identified in chromosome images. Then multi-spectral M-FISH chromosome imaging is presented and some of the possible advantages and difficulties that it might bring to the analysis of chromosome images are discussed.

Section 2.2 introduces chromosomes and chromosome imaging. Section 2.3 describes how chromosomes can be classified. Section 2.4 gives examples of abnormalities that occur in chromosome images. Section 2.5 explains the basic image analysis problems in chromosome imaging and describes how the problems relate to M-FISH imaging. Section 2.6 introduces M-FISH chromosome imaging and the improvements that it offers over traditional grayscale chromosome imaging methods. Section 2.7 explores the possibility of applying image analysis techniques to M-FISH images. Finally, Section 2.8 summarizes the concepts discussed in the chapter.

2.2 CHROMOSOMES

Chromosomes are the body's information carriers. They are the structures that contain genes, which store in strings of DNA all of the data necessary for an organism's development and maintenance. Chromosomes serve as an intricate

blueprint or schematic for cells and organisms. They are found in every nucleated cell in all living organisms. They contain vast amounts of information; in fact, every cell in a normal human being contains 46 chromosomes, which among them have 6×10^9 bits of information [2].

By looking at images of sets of chromosomes in a person, one can collect information about the genetic health of that individual and diagnose certain diseases in that individual. Chromosomes can only be examined visually, however, when they replicate in a process known as mitosis. Under normal circumstances, chromosomes are extremely long and thin and are essentially

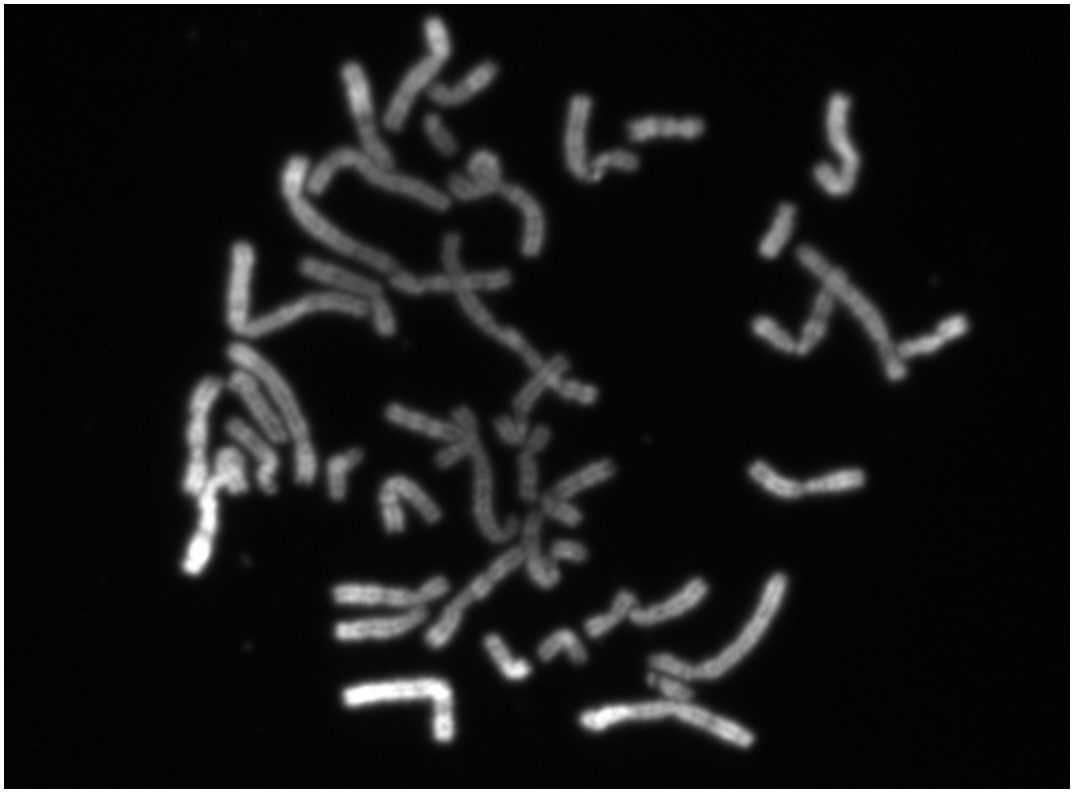


Figure 2.1: Typical chromosome image

invisible. However, during the metaphase stage of mitosis, they contract and become much shorter (around 2-10 μm) and wider (around 1-2 μm diameter). At this stage, they can be stained to become visible and can be imaged by a microscope (see Figure 2.1). The cells for producing these images are commonly obtained from blood specimens, bone marrow, and amniotic fluid.

2.3 KARYOTYPING

Karyotyping is the process of classifying each chromosome in a cell according to a standard nomenclature. In humans, the 46 chromosomes consist of 23 pairs of chromosomes, one of each pair coming from the father and the other from the mother. Of the 46 chromosomes, there are 22 homologous pairs and two sex chromosomes denoted X and Y (see Figure 2.2 and Figure 2.3). A normal human female has two X chromosomes, while a normal male has an X and a Y chromosome. By convention, the 22 pairs and the X chromosome and Y chromosome are assigned to 24 distinct classes, where the first 22 classes are numbered in order of decreasing length (that is, class number one is the longest homologous pair of chromosomes), and the last two classes are for the X and Y chromosomes.

There are several features of chromosomes that have traditionally been used for classification. The first and most obvious of these features is size. Table 2.1 on page 12 shows the average size of each chromosome class in the ADIR (Advanced Digital Imaging Research) M-FISH database, a chromosome image dataset discussed in Section 5.2. Each size is given as the expected percentage of total chromosome area in an image that a chromosome of that class would cover.

Since there are generally two of each chromosome, the sum of these numbers should be approximately one half (the discrepancy being accounted for by the X and Y chromosomes, which do not necessarily occur twice in every normal image).

The second feature traditionally used in karyotyping is the relative centromere position. The centromere is the narrow "neck"-like region in each chromosome. If the centromere is near the middle of a chromosome, that chromosome is said to be metacentric; if the centromere is near the end of chromosome, the chromosome is said to be submetacentric; and if the centromere is at the end of the chromosome, the chromosome is said to be acrocentric.

However, using only the chromosome length and relative position of the centromere, each chromosome cannot be reliably classified into the complete 24 classes of chromosomes, but only one of seven groups known as the Denver classifications [3] (see Figure 2.3). Group A of the Denver classifications includes classes 1-3, the longest chromosomes, which are all metacentric. Group B includes classes 4 and 5, which are long and submetacentric. Group C chromosomes, classes 6-12, are medium sized and submetacentric. Group D, classes 13-15, are moderately short and acrocentric. Group E contains classes 16-18, which are also moderately short, but submetacentric; and groups F (classes 19 and 20) and G (classes 21 and 22) are very short and metacentric and acrocentric, respectively. The X and Y chromosomes are generally classified in Groups C and G, respectively, although they are shown separately in the karyotype representation (see Figure 2.3).

Figure 2.2: Giemsa banded chromosomes

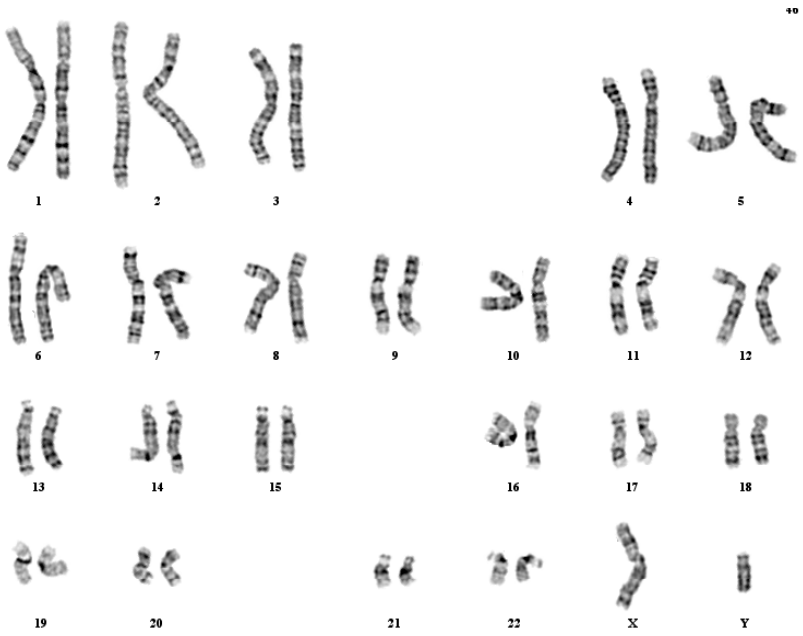


Figure 2.3: Karyotype of Giemsa-banded chromosomes in Figure 2.2

Table 2.1: Average chromosome sizes by class (expressed as the fraction of total chromosome area in an image)

Chromosome Class	***Average Size***	***Denver Group***
1	0.0412	A
2	0.0394	A
3	0.0336	A
4	0.0316	B
5	0.0300	B
6	0.0285	C
7	0.0261	C
8	0.0236	C
9	0.0221	C
10	0.0219	C
11	0.0220	C
12	0.0222	C
13	0.0195	D
14	0.0173	D
15	0.0180	D
16	0.0148	E
17	0.0134	E
18	0.0137	E
19	0.0100	F
20	0.0108	F
21	0.0084	G
22	0.0087	G
X	0.0255	(C)
Y	0.0108	(G)

In order to identify correctly all 24 chromosome types in normal grayscale chromosome images, a banding technique can be used. With proper staining techniques, such as Giemsa banding techniques [4, 5], a unique banding pattern appears on each chromosome type so that all 22 pairs of chromosomes and the X and Y chromosomes can be uniquely identified (see Figure 2.2).

Once all of the chromosomes in a cell have been classified, they can be placed into a graphical representation in which they appear in increasing order of their type number. This representation is known as a karyotype (see Figure 2.3).

2.4 CHROMOSOME ABNORMALITIES

Once the chromosomes have been segmented, one can look for abnormalities in the chromosomes. The most obvious abnormality is an unusual number of chromosomes. Having only one of a type of chromosome is a monosomy, such as Turner's syndrome, in which there is only one X chromosome and no Y. Having three of a type is a trisomy, such as Down's syndrome, in which there are three type 21 chromosomes. It can result in serious mental and developmental retardation.

Other possible abnormalities include deletions. In a deletion, part of a chromosome is lost. An example of a problem caused by a deletion is William's syndrome, a disorder of the circulatory system. In William's syndrome, the gene that produces a protein that affects elasticity in blood vessels is deleted from a type 7 chromosome.

There can also be duplications of genetic material within a chromosome and translocations where two chromosomes exchange genetic information. The

Philadelphia chromosome results from a translocation in the ninth and twenty-second chromosomes. This is often associated with chronic myelogenous leukemia. [6]

In addition, there are a wide variety of other disorders including ring chromosomes, inversions, broken chromosomes, and combinations and variations of the above abnormalities [7]. Detecting these abnormalities is vital because they are reliable indicators of genetic disease and damage and because studying them can lead to new insight about the diseases with which they are correlated. Chromosome abnormalities are particularly useful in cancer diagnosis and research [8].

2.5 Analysis of Grayscale Chromosome Images

Researchers have been studying how best to use computers to aid in chromosome imaging and analysis for over thirty years [9, 10]. These studies, and object recognition problems in general, have traditionally fallen into one of two categories - segmentation or classification. Although there are studies that combine the two categories [11, 12, 13, 14], most publications fall into one or the other of these two categories, and thus they are divided into these categories below.

2.5.1 Segmentation

Segmentation is the process of dividing the image into sections, or segments, each of which has some meaning to a human observer. In chromosome analysis, it is desired to segment the image into background and chromosome pixels, and to divide further the chromosome pixels into individual chromosome

type pixels. Segmenting a chromosome image into background and chromosome pixels is a fairly straightforward task and is usually accomplished with either thresholding or adaptive thresholding. However, dividing the chromosome pixels into individual chromosomes is far from trivial because chromosomes often touch or overlap. At the point of overlap, pixels belong to multiple chromosomes.

Stated more formally, the problem of chromosome segmentation is a problem of partitioning the image into minsets [15, 16]. A minset can be defined as

$$M_{\delta_1,\ldots,\delta_{|K|}} \equiv \bigcap_{i=1}^{|K|} \hat{A}_i; \qquad \hat{A}_i \equiv \begin{cases} A_i' & \text{if } \delta_i = 0 \\ A_i & \text{if } \delta_i = 1 \end{cases} \tag{2.1}$$

where $\{A_i\}_{i\in K}$ is a set of subsets and $\delta_i \in \{0,1\}$. Thus the set of $\delta_1,\ldots,\delta_{|K|}$ is just a binary representation of the minset. Conversely, each subset A_i can be defined by its minsets

$$A_i \equiv \bigcup_{j\in L_i} M_{\delta_1^j,\ldots,\delta_{|K|}^j} \tag{2.2}$$

where L_i is the set of required minsets. For the case of chromosomes, $M_{0,\ldots,0}$ is defined to be the background and every other minset is defined to be a part unique to one chromosome or common to several chromosomes. In the case of touching chromosomes, each chromosome consists of only one minset, while in overlapping chromosomes, each chromosome may be composed of several minsets.

Given an image A containing r objects $\{O_i\}_{i=1}^{r}$, an ideal thresholding operation produces a binary image of objects given by

$$O = \bigcup_{i=1}^{r} O_i \tag{2.3}$$

and background which is the complement of O

$$B = A - O = O' \tag{2.4}$$

Of course, no segmentation is ideal, so what initial segmentation gives is a set of q objects $\{O_i^*\}_{i=1}^{q}$ and an estimated background B^*. For each subset O^*, if it can be partitioned into minsets of $\{O_j\}_{j\in\overline{K}_i}$, where $\overline{K}_i$ is given by

$$\overline{K}_i = \left\{ j \middle| O_j \subset O_i^*, 1 \le j \le r \right\} \tag{2.5}$$

then each object (chromosome) can be written as a union of these minsets.

Agam and Dinstein [16] introduced and effectively used minsets to decompose touching and overlapping chromosomes in the grayscale case. However, it should be noted that any attempt at chromosome segmentation essentially performs minset partitioning and can be written in the minset framework. In their work, they determined minsets using hypothesis testing to choose cut points for dividing clusters of chromosomes. The hypotheses were verified with a chromosome shape model, in which a chromosome is modeled as a rectangle with a contraction (centromere) and at most one bend. A hypothesis was verified if a bounding polygon matching this model could be found that fit closely to the hypothesized chromosome. This method was shown to be successful in many cases but was limited to grayscale chromosome images.

A wide variety of other approaches to the chromosome segmentation problem have also been proposed. A split and merge technique [17] was

proposed that uses a "watershed" [18] method to oversegment the image. A set of seeds are grown using "fall-sets" [19] to determine an initial set of segments. Then adjacent segments in the oversegmented image are rejoined based on a likelihood that is a function of the separating gray level and the common boundary length between the two segments. This is somewhat similar to region-growing [20, 21] methods, which grow seed segments until they meet, combining segments only if they satisfy a certain criteria, such as convexity. However, these methods are only useful for decomposing touching chromosomes and do not handle overlaps.

Fuzzy set theory [22] has also been applied to chromosome segmentation. In this work, a fuzzy binary relation is defined on the boundary points of high curvature, and fuzzy subsets are defined over the points that make up the boundaries between the elements in the cluster. Decompositions are chosen based on the fuzzy relations and the convexity of the resulting segments. However, this method works only for simple cases and fails in the cases of bent chromosomes or clusters of several chromosomes [16].

Valley searching attempts to find a "valley" of gray values that represent a separation between two chromosomes. Vossepoel [2] defines a set of rules and parameters with which to find candidate cut points. It then attempts to link these points with a minimum-cost algorithm that searches for a "valley" connecting these two cut points. This method often works well at finding accurate boundaries, but it also does not handle overlaps.

A model-based method was described in [23] that characterized several different types of boundary features that were shown to be highly correlated with touches and overlaps. It also described relationships between sets of these features that typically occur in touches and overlap. A set of rules was defined which used these boundary features to find possible cut points and group them together to define touches and overlaps. This method showed success in recognizing clusters but had a relatively high failure rate for finding plausible separation paths.

Ji [24] used the concepts of skeletons [25] and convex hulls to decompose overlaps. Overlaps were identified by finding crosses in the skeletons. Candidate cut points were then searched for near this cross. Cut points were chosen based on the curvature of the boundary at that point and on the distance of that point from convex hull of the object. This work has shown some of the most successful results in the literature, but still is limited since it only uses grayscale and geometric information.

2.5.2 Classification

Generally, after segmentation, the next step in chromosome image analysis is classification of the segmented chromosomes. Once the chromosomes have been properly segmented and classified, it is simple to arrange the chromosomes into a karyotype (see Figure 2.3) for examination. After chromosomes have been segmented, chromosomes have a number of features, including length, centromere index, and banding pattern, that can be used to classify them. Length is simple to measure for properly segmented chromosomes,

but centromeres can be subtle and are sometimes difficult to locate. Furthermore, I have already mentioned in Section 2.3 that length and centromere index by themselves cannot be used to classify chromosomes reliably into their 24 classes. For this reason, since its introduction, the banding pattern has been the most popular feature for both manual and automated chromosome classification. However, the difficulty with banding patterns for automated classification is that they often can be difficult to extract. Traditionally a medial axis transform [26] is performed on the chromosomes to straighten it, and then a density profile is measured by integrating the intensities along sections perpendicular to the medial axis of the chromosomes. This could be challenging though, since many bent chromosomes are not trivial to straighten. In addition, overlapped chromosomes were also problematic because part of their banding pattern is not visible at all.

Despite these difficulties, several classification schemes have been developed with some success. Most classification systems use some combination of these same three features, length, centromere index, and banding pattern; but there are many different ways to represent these features.

Several transforms have been proposed for representing chromsome banding patterns. Fourier descriptors have been used as a global description of the chromosome's density profile, and the first eight components of the Fourier transform were found to be most useful for discrimination [27].

Another transform was proposed in [28] which described a set of weighted density distribution functions. These serve as a set of basis functions. Each chromosomes's density profile was correlated with these functions, and the

correlations served as a representation of that chromosome, rather than the profile itself. *This is presently the most commonly used technique for banding pattern classification.*

Laplace local band descriptors [29] have been used to extract only the most dominant bands, since these bands were believed to be the most significant for classification. A two-dimensional Laplacian filter and a set of thresholds were used to determine the size and position of the larger, darker bands on the chromosome. Features from these bands, such as width, position, and average density, are then fed to a classifier.

Markov chains [30] have also been used to represent the banding patterns of chromosomes. In this approach the density profile is quantized, and represented as a chain of symbols. A set of these density profiles from the same class are then used by an inference technique to build a constrained-first order Markov chain that represents this class. When a chromosome is classified, its profile is assigned to the class represented by the Markov chain that is most likely to produce that profile.

A number of different classifiers have been used as well. These include neural networks [31, 32]. In one neural network implementation [33], a multi-layer perceptron neural network was used. Chromosome length, centromere index, and a 15 points from a 64-element density profile, were used as features.

Homologue matching [34, 35] uses two criteria for classifying chromosomes. First for a chromosome to be classified as a certain class, it must be similar to a typical chromosome of that class. Second it must be similar to the

other chromosome of that class within the same image. This is particularly useful for detecting chromosome abnormalities in an image.

Other approaches include fuzzy classifiers [36] whose output is a numerical measure of similarity to a known class and several other statistical methods. A good review of these methods is given in [37].

While considerable success has been achieved with these methods, they all suffer from the same drawback, that they rely on features, such as centromere position and banding pattern, which can be difficult to measure and depend on segmentation accuracy.

2.5.3 Joint Segmentation and Classification Methods

Traditional image analysis methods have viewed segmentation and classification as separate processes. However, the two processes are closely related. Each can be improved with information that the other provides. In the case of chromosome segmentation, this has been realized and suggested before. Ji [38 (see pages 188-189)] recognized the dilemma that classification needs correct information from segmentation, but that segmentation often needs correct information from classification as well; Ji suggested a system of feedback between the two steps. Agam and Dinstein [16] also realized this and suggested combining the two steps for more accurate identification. Both [16] and [38] recognized the potential usefulness of combining segmentation and classification, but provided no method to accomplish it.

Martin [13] demonstrated limited success by combining segmentation and classification in another area of non-rigid object recognition, specifically optical

character recognition. More recently credibility networks were proposed as a framework for joint segmentation and classification [14]. Both of these attempts show promise, but neither is easily extensible to objects without definite shape, such as chromosomes. The first attempt at combining classification and segmentation for chromosome images was introduced in [12], where a classification-driven segmentation method [11] was extended to handle chromosome cluster decomposition, although it was limited to grayscale chromosome images and did not consider overlaps. In addition, it made no provisions for images of multiple clusters or clusters of more than two chromosomes.

The joint chromosome segmentation and classification method developed in this dissertation for multi-spectral images is somewhat similar to the approach for grayscale images in [12] in that it employs a form of classification-driven segmentation. The proposed method uses a set of likelihood functions to accomplish both segmentation and classification simultaneously. It does not employ segmentation-classification feedback and hence does not suffer from error propagation due to outliers in segmentation or classification. Further, the general probabilistic modeling results in an intuitive and extensible framework for the segmentation and classification of chromosomes.

2.5.4 Comparison

Table 2.2 shows a comparison of several different segmentation and classification algorithms. This table must be viewed with a bit of caution. *It is difficult to compare methods directly since published methods rarely use rates*

directly comparable with other work. In segmentation, some work measures touch and decomposition and overlap decomposition separately; other do not distinguish between the two. Some use full real-world images of chromosomes, while other were only tested on a set of pairs of overlapping or touching chromosomes. Furthermore, many published segmentation rates are run on different sets of data. There are at least five grayscale chromosome datasets used in the literature: Copenhagen [29, 37], Edinburgh [33, 37], Philadelphia [37], Delft [29], and Soroka5 [33]. Because of these difficulties in comparing methods directly, I have quantized the published segmentation accuracy rates of the methods to low (<70%), medium (70-80%), and high (>80%).

Classification rates are also difficult to compare directly. Some classify only within a set of chromosome types rather than all 24 types. Some assume perfect segmentation; others do not. In addition, some classification work proposes a new feature representation, while some classification work proposed a new classifier; it is difficult to directly compare the merit of two features if they are not used with the same classifier.

In spite of these difficulties, I have included Table 2.2, as a rough comparison of methods that have published some segmentation and/or classification accuracy rates.

Table 2.2: Segmentation and classification method comparison

Method	***Segmentation Accuracy***		***Classification Rate***	***Joint Segmentation-Classification***
	Touch	***Occlusion***		
Vossepoel [2]	Medium	n/a	n/a	no
Lerner [12, 33]	High	n/a	84%	yes
Agam [16]	High	High	n/a	no
Wu [23]	Low	Low	n/a	no
Li [24, 38]	High	High	n/a	no
Granum [28]	n/a	n/a	90%	no
Groen [29]	n/a	n/a	89%	no
Stanley [35]	n/a	n/a	80%	no

2.6 M-FISH IMAGES

A new way to acquire chromosome images came about with the invention of chromosome painting [39], and combinatorial [40] and ratio labeling [41]. These techniques make use of fluorophores (dyes) that attach to a single type of chromosomes, parts of chromosomes, or specific sequences of DNA. Using these techniques, it is possible to create a combination of fluorophores such that each class of chromosomes absorbs a different combination of these fluorophores [1, 42, 43]. Since each fluorophores has a different emission spectrum, each chromosome class appears to be a different color and is visually distinguishable from all other classes without the aid of banding patterns. An image of each

fluorophore can be obtained by employing appropriate optical filters. With five fluorophores, a multi-spectral image can be obtained in which each pixel is represented as a five-dimensional vector, with each element in the vector representing the magnitude of one fluorophore at that point. Instead of the grayscale image that was obtained in traditional chromosome imaging techniques, a multi-spectral image is now available in which the spectral composition at each point reveals the combination of fluorophores and, thus, the chromosonal origin of the matter at that point. Using this combinatorial labeling, known as M-FISH, it is possible to determine the most likely chromosomal origin at every point in the image [44]. An example of an M-FISH image is shown in Figure 2.4.

Such an imaging technique has a few obvious advantages. First, the task of chromosome classification is greatly simplified. Instead of having to estimate features such as centromere positions and banding patterns, which may be difficult to measure, one only has to look at the spectral information within that chromosome. The second advantage is that it is possible to detect smaller translocations and rearrangements than were discernible with banding patterns only [45]. Small translocations are easily noticed as a single chromosome with two different colors in it.

With M-FISH images, an entirely new source of information is available for segmentation as well. If one observes the example in Figure 2.5, it is not immediately clear, by looking only at the boundary of the cluster, what the proper segmentation of the cluster is. It is not apparent, even to many human observers, whether there is an overlap involved or even how many chromosomes are

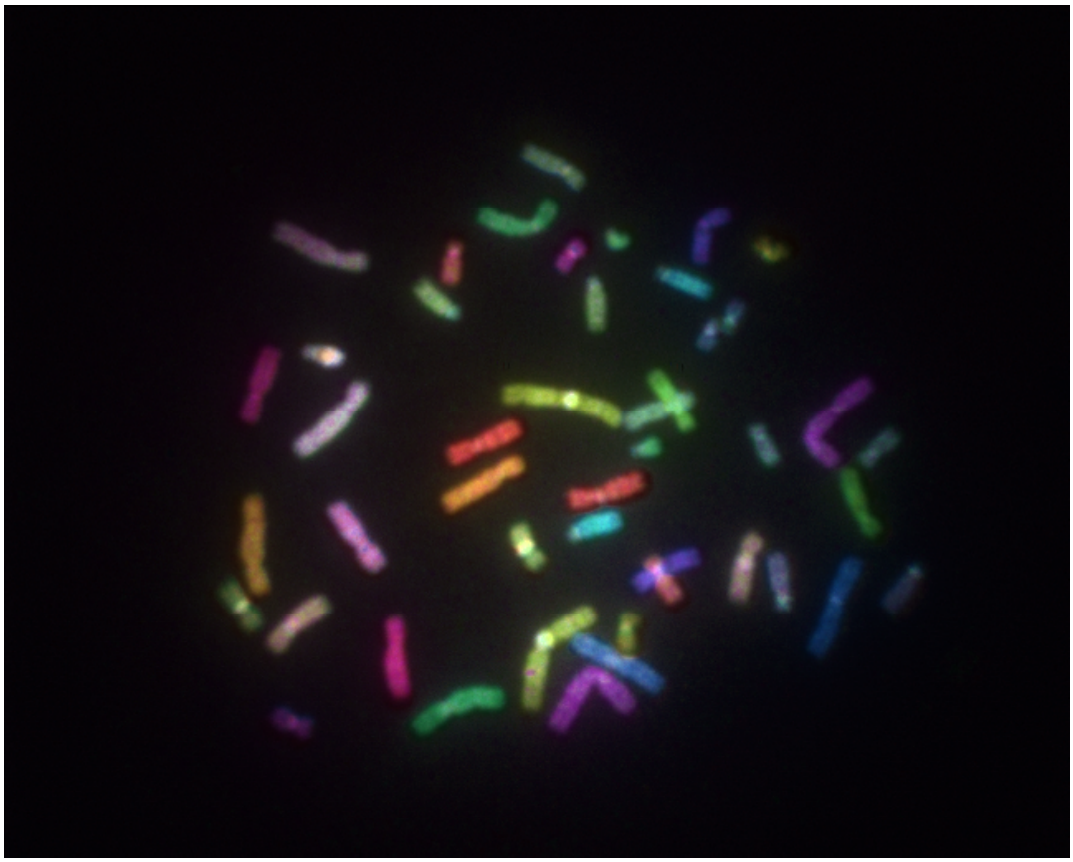
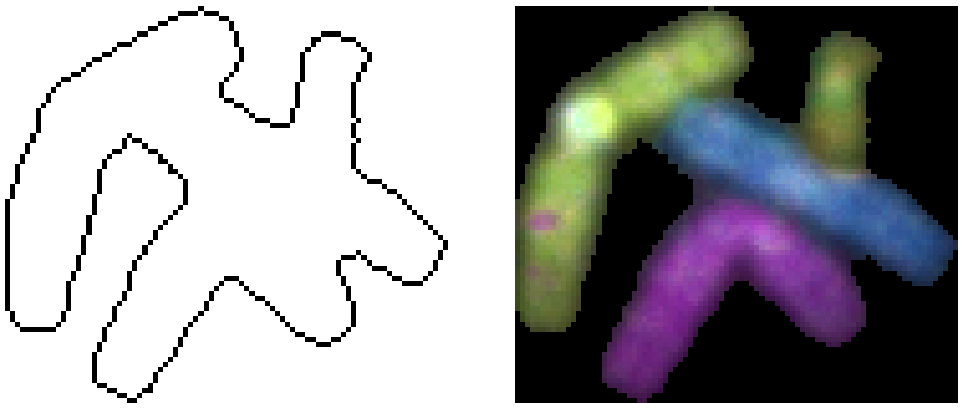


Figure 2.4: M-FISH image

included in this cluster. However, by looking at the M-FISH multi-spectral information, a human observer would very easily be able to determine what the proper segmentation should be since each chromosome has its own color.

Several sets of fluorophores are commonly used for M-FISH imaging. In all these sets, one fluorophore, DAPI (4',6-Diamidino-2-phenylindole), which attaches to DNA and thus labels all chromosomes, is typically used to generate a traditional grayscale image of the chromosomes (see Figure 2.1). Five additional fluorophores are used to distinguish class. The combinations of these five

fluorophores that are used to label each type are shown in the Appendix for three different M-FISH fluorophore sets. However, these tables are somewhat of an oversimplification because, in practice, fluorophore absorption is hardly binary. Table 2.3 shows the actual mean values of pixels of each class from a real set of M-FISH images. The predicted absorbed fluorophores are boldfaced to prevent the reader from having to cross-reference with the table in the Appendix. As one can see, the strength of absorption is not binary and varies widely across the chart. Both class 20 and class 3 are predicted to absorb Spectrum Orange, but Spectrum Orange is almost twice as strong in class 20. Also in this particular image set, the Cy5.5 fluorophore is weak; and its strength in classes that should absorb it is occasionally less than that of other dyes in classes that should not. Furthermore, the difference in magnitude of classes that should absorb Cy5.5 and classes that should not is not always great. The average magnitudes of Cy5.5 in classes 4 and 5 are nearly identical, although class 5 should bind Cy5.5, while class 4 should not. In addition, it is important to note that the characteristics in this table are valid only for this set of data, since fluorophore strength often varies by batch and by age of the fluorophore.



(a) Boundary of cluster

(b) Multi-spectral information in cluster

Figure 2.5: Comparison of two types of cluster information

Because of these difficulties, a simple threshold for each dye is usually not sufficient for reliable pixel classification, even for a single image. A classifier must be designed that takes advantage of disparities such as that of Spectrum Orange in classes 20 and 3, and compensates for fluorophores, such as Cy5.5 in the image set characterized by Table 2.3, which may not be good differentiators of class. A Bayesian classifier has been proposed in [44] that can be trained on each set of images to compensate for different fluorophore characteristics that may occur in each set. The method calculates the maximum *a posteriori* probability of a pixel belonging to each class, and the most likely class is selected. This classifier has proven to be very successful, and the pixel classifier used in this dissertation is based on this work (see Section 3.3).

Table 2.3: Average fluorophore magnitude for each class in image subset A06 of the ADIR chromosome image dataset. Bold denotes the classes to which each fluorophore is predicted to bind.

Chromosome Class	***Spectrum Green***	***Spectrum Orange***	***Texas Red***	***Cy5***	***Cy5.5***
1	**0.5483**	0.2946	**0.4928**	**0.5171**	0.2554
2	0.4681	0.3596	0.4117	0.4311	**0.4978**
3	**0.5059**	**0.4549**	0.3350	**0.5200**	**0.3434**
4	**0.5852**	0.3447	0.3882	**0.5119**	0.3041
5	**0.5372**	**0.4523**	**0.5299**	0.3077	**0.3114**
6	**0.5390**	0.2577	**0.5019**	**0.4823**	**0.3469**
7	0.3244	0.2560	**0.5794**	**0.6313**	0.2453
8	**0.8027**	0.2842	0.3140	0.3034	0.2304
9	**0.6379**	**0.4764**	0.3160	0.2863	**0.3796**
10	0.3563	0.2809	0.3219	**0.6858**	**0.4257**
11	**0.5913**	**0.4987**	0.2877	**0.4994**	0.2066
12	0.3127	0.2479	**0.7338**	0.3297	**0.3945**
13	**0.6590**	**0.5828**	0.3083	0.2367	0.1849
14	0.3266	0.2946	**0.7695**	0.3396	0.2279
15	0.2590	**0.5066**	**0.6101**	**0.4857**	0.1936
16	**0.6752**	0.2025	**0.6194**	0.2544	0.1698
17	0.3739	0.2928	0.3339	**0.7386**	0.2823
18	**0.6085**	**0.5011**	**0.5353**	0.2151	0.1576
19	0.2917	**0.6466**	0.3369	**0.5539**	0.2019
20	0.2746	**0.8125**	0.3551	0.2596	0.1988
21	**0.5994**	0.3411	0.3636	0.3547	**0.4403**
22	0.2603	**0.4837**	**0.5860**	**0.4697**	**0.3041**
X	0.4014	**0.5829**	0.3966	0.3793	**0.3913**
Y	**0.6486**	0.2267	0.2274	**0.5492**	**0.3632**

During pixel classification, special care must be taken with areas of overlap. Since with M-FISH, the chromosomes are illuminated from above and viewed from above, the major contribution for a pixel in an area of overlap will come from the top chromosome. However, in practice, the chromosomes are somewhat opaque, so that pixel will include information from both chromosomes. This could lead to a pixel being classified as the same type as the top chromosome, the same type as the bottom chromosome, or neither.

2.7 ANALYSIS OF M-FISH IMAGES

All of the chromosome segmentation and classification techniques mentioned previously have been developed for grayscale chromosome images. For 30 years now, many researchers have studied image analysis of grayscale chromosome images. These studies have resulted in significant improvements in techniques for chromosome segmentation and classification, and they have greatly simplified the work of lab technicians and those evaluating karyotypes.

However, to date there is little work on image analysis of M-FISH chromosomes images. An entropy criterion for segmenting class maps of chromosome cluster was explored in [46]. This was an early, primitive attempt at using multi-spectral information to segment chromosome images. Some success was shown, but the success of the method was very sensitive to its parameters, and it was not robust over a wide variety of images. Furthermore, this method only performed chromosome segmentation. No chromosome classification method was proposed, and thus classification information could not be used to aid

in segmentation. The entropy approach was extended to use entropy estimation for application directly to M-FISH data [47], but it achieved little success.

The next step in the evolution of chromosome imaging is the application of image analysis and pattern recognition techniques to the multi-spectral M-FISH images. These images provide significantly more information than grayscale chromosome images and promise significant improvements in the accuracy of chromosome identification, classification, and anomaly detection. While 24-color chromosome labeling [1] has greatly simplified the classification of chromosomes, it is not immediately clear what the best way is to use this multi-spectral method to segment the image and decompose touching and overlapping chromosomes.

The same problem posed in the context of multi-spectral images varies significantly from the grayscale case. In the grayscale case, it was assumed that thresholding, or binary segmentation, would result in as many or fewer objects than there were chromosomes. However, in the multi-spectral case, an initial segmentation can break the image into *at least* as many objects as there are chromosomes in the image. That is, $q \leq r$ in the minset notation (Section 2.5.1). For example, it is likely that two overlapping chromosomes could be segmented initially into three parts, the chromosome on top and the two ends of the chromosome on the bottom. Whereas chromosome segmentation in the grayscale case was a "splitting" problem, we will see that it becomes a "merging" problem in the multi-spectral case.

A further consideration for any useful multi-spectral segmentation technique is that it must be able to resolve touching and overlapping chromosomes without losing the ability to detect translocations and rearrangements. A useful criterion must be found for distinguishing between translocations, in which a chromosome may be made up of two colors, and touching (or overlapping) chromosomes, in which two separate chromosomes of different colors appear to be connected.

As for chromosome classification, M-FISH eliminates many of the prior difficulties encountered in chromosome classification. No longer are centromere location, banding pattern, and other complicated, difficult to measure, features necessary to determine a chromosome's class since color alone is theoretically sufficient to determine the class. Since each pixel can be classified individually, each chromosome is just assigned to the class to which most of its pixels have been classified.

Yet another benefit for M-FISH is that classification can be performed independently of segmentation. Grayscale methods were often forced to perform segmentation followed by classification, since the grayscale classification features, length, centromere index, and banding pattern, could only be measured on a segmented chromosome. With M-FISH images, I can reliably estimate what class a pixel belongs to before I even know what segment it is part of. This is very useful, as we will see in Chapter 3, since I can use this classification information for more accurate segmentation, and this more accurate segmentation will, in turn, give us more accurate classification.

2.8 CONCLUSIONS

This chapter introduced the basic terminology and concepts used in this dissertation. It briefly described traditional grayscale chromosome imaging methods, and a new chromosome imaging method, M-FISH, which generates a multi-spectral image of chromosomes, was introduced. While some degree of success has been achieved with traditional grayscale methods, and while one could just apply these methods to the DAPI channel of M-FISH images, I showed an example in which the correct decomposition of a cluster of chromosomes is made evident only with multi-spectral information. I suggested that this multi-spectral information could be used for improved segmentation of chromosomes because touching and overlapping chromosomes should be easier to resolve in M-FISH images. Whereas chromosome segmentation in the grayscale case was a "splitting" problem, it becomes a "merging" problem in the multi-spectral case. Based on this observation, the following chapters develop a method that shows how improved segmentation and classification can be accomplished by using the multi-spectral information in M-FISH.

Chapter 3: Maximum Likelihood Algorithm for Joint Segmentation-Classification

3.1 INTRODUCTION

This chapter formulates the chromosome segmentation and classification problem as a unified maximum likelihood (ML) hypothesis testing problem. Unlike earlier chromosome analysis approaches which first perform segmentation before classification, the proposed formulation results in a joint segmentation-classification strategy. The likelihood function I use for the hypothesis test utilizes chromosome size and multi-spectral pixel data. It is scale-invariant, so it is not affected by microscope magnification or the stage of mitosis in which the chromosomes were captured; and it accounts for overlapped chromosomes in the model, so it is able to correctly identify chromosomes even if they are partially covered by other chromosomes.

Section 3.2 states the problem of chromosome segmentation and classification as a maximum likelihood hypothesis test. Section 3.3 discusses the mathematical formulation of a robust scale-invariant likelihood function that is able to account for overlapping chromosomes and implicit segmentation errors. This is accomplished by incorporating a weighting function into the likelihood function computation. The hypothesis test presented in this chapter applies to a set of possible chromosome segmentations. Section 3.4 concludes the chapter.

This chapter does not describe how the possible chromosome segmentations are chosen. This issue is the central theme of Chapter 4.

3.2 PROBLEM FORMULATION

In this section, I give a formulation of the chromosome segmentation-classification problem as a maximum likelihood hypothesis test. I use this formulation in the following section to segment and classify multi-spectral chromosome images efficiently.

I define C_i as the set of all pixels belonging to class i. Since there are 24 classes of chromosomes each non-background pixel may be classified as one of 24 classes. I do not explicitly handle background pixels since I assume that background/foreground segmentation is performed as a preprocessing step before any other segmentation is carried out.

A_i^n denotes the set of pixels belonging to the n^{th} chromosome of class i in a single image, or in a set of images. Since, within any set of images, several chromosomes may belong to the same class, $A_i^n \subseteq C_i$. $\left|A_i^n\right|$ denotes the cardinality of the set, which is the number of pixels in the chromosome.

I want to find the sets A_i^n, that represent the chromosomes that need to be segmented and classified. In general, given a likelihood function, or a measure, of the probability that an arbitrary segmented object is a chromosome A_i^n, the segmentation-classification problem reduces to choosing the segmented objects and corresponding classes to maximize the likelihood function. I also need a mechanism for generating candidate chromosomes for evaluation using the maximum likelihood hypothesis test. This is the subject of Chapter 4.

Thus the joint chromosome segmentation-classification problem is composed of three steps:

1. designing a suitable likelihood function for evaluating the likelihood of a given candidate chromosome being a chromosome of a certain class
2. generating sets of candidate chromosomes, and
3. choosing the best set of candidate chromosomes and classes to which they belong from the maximum likelihood test.

3.3 PROPOSED LIKELIHOOD FUNCTION

The proposed likelihood function $L(\cdot)$ is a product of two separate likelihood functions $L_{multi}(\cdot)$ and $L_{size}(\cdot)$ and a weighting function $w(\cdot)$ that accounts for overlaps and improves segmentation accuracy. While the likelihood function $L_{multi}(\cdot)$ uses the multi-spectral information, $L_{size}(\cdot)$ uses information on the relative chromosome size. $L(\cdot)$ is a function of a possible chromosome segmentation and a possible class. Thus $L(\cdot)$ incorporates information central to both segmentation and classification. Since all three components, $L_{multi}(\cdot)$, $L_{size}(\cdot)$, and $w(\cdot)$, must be between 0 and 1, the entire likelihood function must also be between 0 and 1. I choose the product, rather than the average or a weighted combination, to combine the three components, so that the value of all three components must be large in order to make the total likelihood value large. In the following description, I refer to a possible segmentation of a single chromosome as a candidate chromosome.

Definition 1. Given a candidate chromosome A', the likelihood that A' belongs to the class i, due to the multi-spectral data in its pixels $\mathbf{x}$, is given by $L_{multi}(A', i) = \mathrm{E}\left[p(\mathbf{m} \in C_i | \mathbf{x}(\mathbf{m})) | \mathbf{m} \in A'\right]$.

This likelihood measure $L_{multi}(\cdot)$ in Definition 1 is the average of the probabilities of each of the pixels in A' belonging to class i. Therefore, objects that have a large number of pixels with a high probability of belonging to class i will result in a large likelihood function for class i. From [44], I use Bayes' theorem to calculate $p(C_i|\mathbf{x})$ as

$$p(C_i|\mathbf{x}) = \frac{p(\mathbf{x}|C_i)p(C_i)}{p(\mathbf{x})} \tag{3.1}$$

In (3.1), I estimate the terms $p(\mathbf{x}|C_i)$, $p(\mathbf{x})$, and $p(C_i)$ from training data by fitting a Gaussian Mixture Model to determine the conditional distributions. These terms can be calculated as follows:

$$p(\mathbf{x}|C_i) = G(\mathbf{x}, \boldsymbol{\mu}_{1,i}, \boldsymbol{\Sigma}_{1,i}) \tag{3.2}$$

$$p(\mathbf{x}) = \sum_i p(\mathbf{x}|C_i) \tag{3.3}$$

$$p(C_i) = \frac{\sum_n \left|A_i^n\right|}{\sum_n \sum_j \left|A_j^n\right|} \tag{3.4}$$

Recall that $G(\cdot,\cdot,\cdot)$ is a Gaussian probability density function (Section 1.3). The means and covariance matrices ($\boldsymbol{\mu}_{1,i}$ and $\boldsymbol{\Sigma}_{1,i}$, respectively) are computed

using maximum likelihood parameter estimation [48] on the training set. The subscript 1 in $\boldsymbol{\mu}_{1,i}$ and $\boldsymbol{\Sigma}_{1,i}$ is used to distinguish these means and variances from those in Definition 2, and the subscript i denotes class. In the case of M-FISH images, training should be applied to each batch because each batch has its own set of dye characteristics (see Section 2.6). Training can be accomplished by using a few images that have been hand segmented.

The prior class probabilities, $p(C_i)$, also must be computed by training on a set of data. However, since relative chromosome size does not vary from image to image, this does not require retraining. For this work, I use the sizes given in Table 2.1, to calculate the prior class probabilities. The only computation necessary to convert the values in this table into prior class probabilities is to multiply the values for the first 22 classes by 2 since each image generally contains two of each class of chromosome. The values for the X and Y chromosomes can be used directly from the table since there might be only one of each of them in an image. Some might argue that an X chromosome is more likely than a Y chromosome since a female karyotype (XX) is as likely as a male (XY). If one assumes male and female karyotypes equally likely, there would be, on average, one and a half X chromosomes and one half Y chromosomes per image, but I have ignored that detail for simplicity. Accounting for this makes very little difference in the overall outcome, and in practice, it is often known ahead of time whether the karyotype is male or female, so no distribution of male and female karyotypes would need to be assumed.

As a complement to multi-spectral information, I also define another likelihood measure to avoid the erroneous segmentation of chromosomes into small segments of locally similar pixels.

Definition 2. Given a candidate chromosome A', the likelihood that A' belongs to the class i, due to its size, is given by $L_{size}(A', i) = G\left(\frac{|A'|}{y}, \mu_{2,i}, \sigma_{2,i}\right)$ where $y = \sum_n \sum_j \left|A_j^n\right|$.

This second likelihood is a function of object size. If the size of the candidate chromosome A' is equal to $\mu_{2,i}$, the mean size of class i, then the likelihood will be greatest. As the size moves farther away from the mean size of class i, the likelihood will become lower. The values used for $\mu_{2,i}$ in this work are shown in Table 2.1. The size variance of each class i is denoted by $\sigma_{2,i}$.

The size of a chromosome used in this function is its relative size, or the percentage of total chromosome area in the image that a chromosome covers. This makes the likelihood function scale invariant. Hence, while a change in microscope magnification might produce larger chromosomes, it would result in the same value for the likelihood function due to the normalization of the chromosome size by y, which is the total chromosome area within the image.

Using this second likelihood function accomplishes several things. First it adds a second completely different source of information useful in classifying chromosomes for more accurate overall classification. As mentioned in Section

2.3, size alone is not sufficient to classify a chromosome reliably into one of the 24 classes. However, it can serve as a check for the first likelihood. If the first likelihood function gave similar likelihood values for a candidate chromosome for classes 1 and 22 (the largest and smallest chromosomes, selected just for the point of example), the likelihood based on size would certainly help distinguish between the two.

The likelihood function in Definition 2 also distinguishes fragments from whole chromosomes. Without it, an oversegmented chromosome would be no less likely than a correctly segmented chromosome, since both would have the same multi-spectral information, and thus the same value for $L_{multi}(A')$, and a broken chromosome or a section of a translocation would be indistinguishable from a normal chromosome. In addition, a likelihood based on size is very useful for detecting clusters of chromosomes, since a cluster of chromosomes will generally be larger than any of the classes given high likelihood values by $L_{multi}(A')$.

In addition to these two likelihood functions, one final component is defined to model overlaps.

Definition 3. Given a candidate chromosome A', the certainty, $w(A')$, of the likelihood functions described in Definitions 1 and 2 is defined to be the percentage of visible, or non-overlapped, pixels in the candidate chromosome A'.

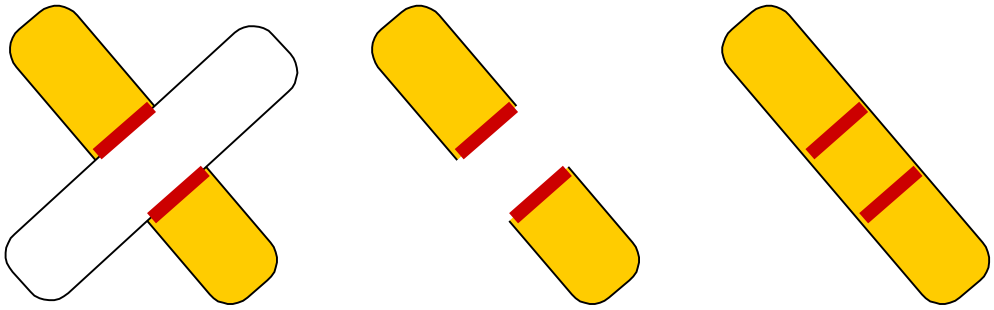
(a) Incorrectly segmented b) Correctly segmented

Figure 3.1: Shaded areas represent two possible segmentations, A', of a single chromosome. The function $w(A')$ gives more weight to case b).

Therefore chromosomes that are overlapped will be less certain than chromosomes that are completely visible, since the function has less information about them. $w(A')$ acts as a weighting function of the overall likelihood function that follows. This may be viewed as an adjustment to take into account overlapping chromosomes, since the weighting function returns a value of unity when there is no possible overlap.

Incorporating the weighting function from Definition 3 also improves segmentation accuracy by avoiding segments being left out from the middle of chromosomes. Figure 3.1 illustrates this possibility. Since A' in Figure 3.1(a) consists of two connected components, the middle (white) portion of the chromosome is assumed to be an area of overlap and the size of both segmentations, and thus $L_{size}(A')$, is calculated to be the same. Without the weighting function, $w(A')$, if $L_{multi}(A')$ were equal for both segmentations, $L(A')$ would also then be equal for both segmentations, although Figure 3.1(b) is the correct segmentation and should receive a higher likelihood.

For the purposes of the weighting function, A' is given as the non-overlapped area of the candidate chromosome. The overlapped area can then be estimated as the area between the connected components of A' (see Figure 3.2).



(a) Cluster of overlapping chromosomes

(b) A', the ends of overlapped chromosome

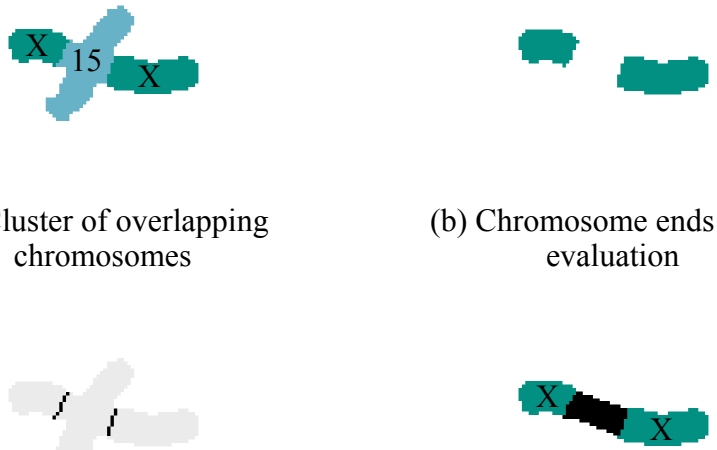
(c) Overlapped area filled in

Figure 3.2: Calculating the weighting function for an overlapped chromosome

If A' contains only one connected component, it is assumed that the candidate chromosome is not overlapped, and thus, its weighting function $w(A')$ is unity.

I estimate the area of overlap from A' as follows. I first take the set of all pixels in the two segments that border the rest of the connected component. Next I draw a line in between each pixel in the first set of border pixels to every pixel in the second set of border pixels. This does not necessarily guarantee a continuous area without gaps, so I finish by filling in all the holes in this new segment. This new segment with the holes filled estimates the overlapped area of the chromosome. This overlapped segment is not used in calculating the probability from pixel data, but it is necessary in calculating the overall size of the chromosome for the $L_{size}(A')$ likelihood function and for the percentage of visible pixels used in the $w(A')$ weighting function.

Figure 3.3 shows an example of the estimation of the overlapped area with an actual chromosome. In this example, a type 15 chromosome is overlapping an X chromosome. Figure 3.3(b) shows the ends of the X chromosome to be



(a) Cluster of overlapping chromosomes

(b) Chromosome ends under evaluation

(c) Border pixels (in black)

(d) Overlapped area estimated (in black)

Figure 3.3: Calculating the area of overlap

evaluated. In order to calculate the area of overlap, I find the pixels that border the rest of the cluster (Figure 3.3(c)), and connect all these border pixels. This gives us the estimate of the overlapped area in Figure 3.3(d). In this case, the estimated area of overlap is 136 pixels, while the visible ends of the X chromosome have an area of 377 pixels, so if A' is the area shown in Figure 3.3(b), the weighting function can be calculated as follows:

$$w(A') = \frac{\text{visible area}}{\text{estimated total area}} = \frac{377}{377+136} = 0.74$$

Finally I combine all three Definitions to obtain an overall likelihood.

Definition 4. Given a candidate chromosome A', the overall likelihood that A' belongs to the class i is given by the product of the likelihood functions and certainty given in Definitions 1 through 3, $L(A',i) = L_{multi}(A',i)L_{size}(A',i)w(A')$.

The classification is then accomplished using the maximum likelihood hypothesis on the candidate chromosome A'. That is, the most likely class is given by the value of i that maximizes the function $L(A',i)$ for a given A'. Similarly, segmentation is accomplished using the maximum likelihood hypothesis on a set of possible segmentations. Whereas classification is maximizing $L(A',i)$ over i, segmentation is maximizing $L(A',i)$ over A'. By maximizing both A' and i over $L(A',i)$, one can simultaneously accomplish segmentation and classification:

$$\arg\max_{A',i} L(A',i) \tag{3.5}$$

In this case, maximum likelihood classification is essentially equivalent to maximum *a posteriori* classification since the prior probabilities for each class are mostly equal. That is, for any candidate chromosome, all classes are equally as likely since there is an equal number of each class of chromosome (2) in most images. The one exception to this is the X and Y chromosomes. Rather than having two of each of them, a normal image will have a pair of X's and no Y's (female), or one of each (male). However, since this exception concerns only 2 of the 24 classes, I have chosen to ignore it and approximate the maximum *a posteriori* classifier with maximum likelihood. The other possibility would be to assume that male and female karyotypes were equally as likely so that there would be 1 Y chromosome and 3 X chromosomes per 2 images, or 92

chromosomes. This would give priors of 1/92 for the Y class, and 3/92 for the X class.

3.4 CONCLUSIONS

This chapter derives a unified maximum likelihood hypothesis test for the joint segmentation and classification of chromosome images. I have shown how the multi-spectral information, chromosome size, and statistical data from training sets may be used to formulate a joint segmentation-classification likelihood function. I have also discussed how to incorporate confidence measures into the likelihood computation to add robustness to the algorithm. The chapter describes how to initialize the parameters used in the maximum likelihood test using multi-spectral chromosome images.

Other likelihood functions could be incorporated in the overall function as well, including possibly a likelihood function based on the shape of the chromosome, or a likelihood function based on other chromosomes within the image (i.e. if another two chromosomes have higher likelihood values of being class 1, then a third chromosome is likely not a class 1 chromosome as well). I have chosen these two likelihood functions based on multi-spectral pixel data and size data in this work because of their ease of implementation and superior performance.

While the likelihood function developed in the chapter is specific to multi-spectral chromosome images, it is also important to note that this framework of unified segmentation-classification is not necessarily constrained to the chromosome segmentation-classification problem. This joint segmentation-

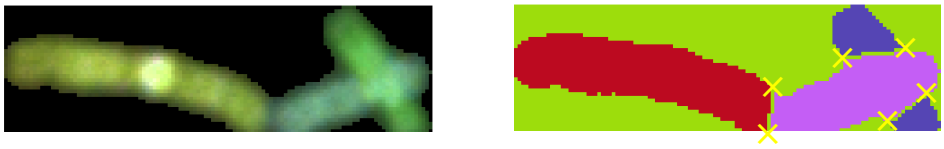
classification framework could just as easily be used on any other segmentation-classification problem given an appropriate likelihood function to measure the quality of both segmentation and classification.

Chapter 4: M-FISH Classification-Segmentation Implementation

4.1 INTRODUCTION

This chapter describes how candidate chromosomes are determined and used in the maximum likelihood hypothesis test to generate a joint segmentation-classification result. Oversegmentation is performed by a Bayesian pixel classifier which is followed by suitable postprocessing to remove small-scale classifier noise. The component segments are then combined pairwise to give candidate chromosomes which are then selected for rejoining using the maximum likelihood hypothesis test. The rejoining is repeated until no suitable pairs of segments can be found. This approach yields efficient and accurate segmentation and classification of M-FISH chromosomes and is able to segment and classify both touching and overlapping chromosomes in one unified approach. A useful byproduct of this approach is the identification of abnormal chromosomes.

It is important to contrast my work with traditional chromosome segmentation methods. Traditional methods began with clusters of chromosomes and attempted to divide them into individual chromosomes by choosing cut points on the boundary of the cluster, which are the points at which the boundaries of the two different chromosomes meet. For the case of two touching chromosomes, two points must be found that define a line that separates the two chromosomes. For the case of two overlapping chromosomes, four cut points must be found that



(a) Color representation of M-FISH cluster

(b) Segmented cluster

Figure 4.1: Typical chromosome cluster in M-FISH image segmented with cutlines. Yellow crosses mark cut points. In this case, lines closely approximate the boundaries between chromosomes.

create a polygon that denotes the area of overlap. Once the proper cut points are discovered, the touching or overlapping chromosomes can then be decomposed by straight cut lines between the points [16] or best fit cubic curves [24] (See Figure 4.1). Whereas traditional approaches such as these have often begun with undersegmented objects, I take the opposite approach and begin by oversegmenting chromosomes and then merging the segments. These segments are derived from multi-spectral information and pixel classification, and these segments, or combinations of them, are often able to represent the more intricate boundaries between chromosomes more accurately than a single cutline (see Figure 4.2).

Traditional chromosome segmentation methods use shape information from the boundary of the chromosomes as a criterion for selecting possible segmentations and for detecting clusters. Methods that search for branches in skeletons [24] have been used to detect clusters. Many algorithms have examined the shape of the boundary of clusters to select cut points [16, 23, 24]. Occasionally, grayscale information from inside the chromosome clusters has also

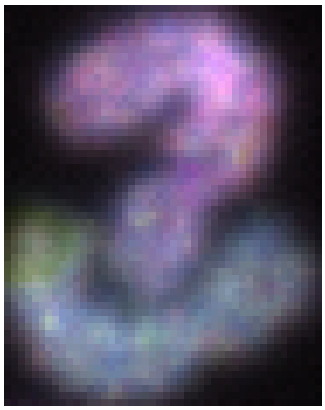


Figure 4.2: Cluster that cannot be split with a cutline

been used. One popular method used "valley searching" [2] where a minimum cost algorithm attempted to locate low gray-value valleys running through the cluster to locate separation between the chromosomes. I choose to use the multi-spectral information available in M-FISH as a criteria for selecting from a set of segmentation possibilities, by incorporating it into a likelihood function to evaluate these possibilities. I then select the most likely via a maximum-likelihood hypothesis test. Chapter 3 describes the maximum likelihood function for multi-spectral chromosome images.

4.2 DETERMINATION OF CANDIDATE CHROMOSOMES

Section 3.3 posed the segmentation-classification problem as a maximum likelihood problem in which I attempted to maximize the likelihood function $L(A', i)$. Since there are only 24 possible classes for a chromosome, it is simple to do an exhaustive search over all possible values of i for any particular candidate chromosome A'. However, there are an extremely large number of possible segmentations for A', so this formulation is only useful if one can

somehow first develop a reasonably limited set of candidate chromosomes from which to choose. Until now, I have assumed that these candidate chromosomes are available for the maximum likelihood hypothesis test. I now describe explicitly how these candidate chromosomes are generated.

Many previous chromosome segmentation methods [38] begin with thresholding, or an adaptive thresholding step. In this work, I do not concern myself with this problem. Instead, I will focus on the problem of decomposing clusters of overlapping and touching chromosomes since single chromosomes will be segmented using only the background/foreground segmentation. Of course, erroneous background/foreground segmentation will lead to erroneous cluster decomposition; however, for the purposes of this work, I will assume that background/foreground segmentation has been performed ideally, or at least close to ideally.

I then perform connected component analysis to parse the image into single chromosomes and clusters of touching and overlapping chromosomes. The result of this processing is a set of r connected components, or objects, $O_1^* \ldots O_r^*$. At this stage, I test each of the objects, O_i^*, using the likelihood function developed in Section 3.3 defined in Definition 4. If O_i^* were a complete chromosome, evaluation of the likelihood function for the correct class would result in a large value; if O_i^* were composed of several touching and/or overlapping chromosomes, evaluation of the likelihood function for any class would result in a small likelihood function value. Note that if O_i^* were a single, but abnormal chromosome, it would also result in a small likelihood function

value. An empirically determined likelihood threshold, T_1, is used to determine if the connected component O_i^* is a single, normal chromosome. If the maximum likelihood function evaluated on O_i^* over all classes is above this threshold T_1, processing of the connected component is terminated since I have both segmented and classified it with a high likelihood.

If the evaluation of the likelihood function results in a value below T_1, then the connected component could either be 1) a cluster of touching and overlapping chromosomes, 2) an abnormal chromosome, such as a broken chromosome or translocation, or 3) a combination of 1 and 2. All of these cases are handled in a unified manner using pixel classification and post-processing the classification map. The post-processing step reduces noise from the pixel classification, improves computational efficiency, and increases segmentation-classification accuracy. The objective of the pixel classification and post-processing is to partition the connected component O_b^* into the mutually disjoint sets $O_{b,1}^* \ldots O_{b,q}^*$, where b indexes a connected component whose likelihood function was evaluated below T_1 for all classes. Since the sets completely make up O_b^*,

$$O_b^* = \bigcup_{j=1}^{q} O_{b,j}^* \tag{4.1}$$

and since the sets are mutually disjoint

$$\bigcap_{j=1}^{q} O_{b,j}^* = \varnothing \; \forall b \tag{4.2}$$

Section 4.2.1 describes the details of this partitioning process, and Section 4.2.2 explains how the sections are merged back together.

4.2.1 Pixel Classification and Post-processing

Each connected component O_b is first processed through a pixel classifier using the maximum *a posteriori* probability $p(C_i|\mathbf{x})$ as discussed in Section 3.3. Since pixel classification is an inherently noisy process, some isolated pixels and small segments could have been misclassified in this step. To reduce noise, I filter the class map using a non-linear majority filtering approach [49]. I choose the majority filter because it removes small segments, but maintains the shape and position of large-scale edges. A majority filter consists of a structuring element H. The image is scanned in raster order, and the class at the center pixel location is replaced by the majority class within the spatial extent of the structuring element. Mathematically,

$$y(\mathbf{m}) = \underset{\mathbf{k}\in H,(\mathbf{m}-\mathbf{k})\in\{O_i^*\}}{\text{maj}} \{x(\mathbf{m}-\mathbf{k})\} \tag{4.3}$$

where x is the input class map, y is the output class map, and *maj* denotes the majority operation. Notice that only object pixels are used for calculating the majority, not background pixels. For my implementation, I use a fairly large structuring element H defined by $H = \{(-8,-8),(-8,-7),\ldots,(8,8)\}$ which represents a 17 × 17 square intended for use with 517 × 645 images. This structuring element should be as large as possible to remove the most noise; however, it cannot be so large that it would remove small chromosomes. I have chosen this structuring element to be about the same size as the smallest

chromosome in an average image. The proposed structuring element was not large enough to filter out even the smallest chromosome in the ADIR M-FISH chromosome image database (see Section 5.2). Another possibility would be to vary the size of the structuring element for each image by looking at the overall chromosome area in the image and then selecting a structuring element that is smaller the expected size of the smallest chromosome for that image. It should be noted that a large majority filter might also remove small translocations. This is acceptable though, since one does not necessarily want to split translocations into two segments. Instead, the object will be identified as a translocation by its low likelihood value.

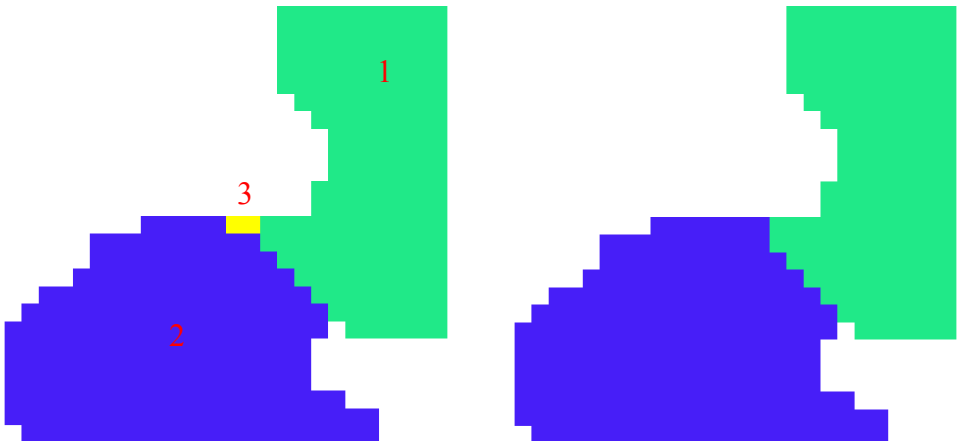
I follow the majority filtering by reclassifying small segments to the most likely class of one of their neighboring segments. This eliminates any remaining small segments. If S_j is the set of pixels in the segment under examination, I define S_j to be a small segment if $|S_j|$ is less than a given threshold, T_2. For each small segment, I call the set of classes of the adjacent segments D_{S_j}. I say that two segments, S_{j_1} and S_{j_2}, are adjacent if

$$\exists \mathbf{x}(\mathbf{m}) \in S_{j_2} \text{ such that } \mathbf{x}(\mathbf{m} - \nabla) \in S_{j_1} \tag{4.4}$$

where ∇ is given by the four-connected set $\{(0,1),(1,0),(0,-1),(-1,0)\}$. The most likely class, $\hat{i}$, is determined by selecting the most likely class for S_j from among only the classes of its neighboring segments, D_{S_j}

$$\hat{i} = \max_{i \in D_{s_i}} L_1(S_j, i) \tag{4.5}$$

where $L_1(\cdot)$ is given in Definition 1 on page 37.



(a) Small segment is #3 (b) Reclassified as neighbor class

Figure 4.3: Small segment reclassification

Figure 4.3 shows an example of small segment reclassification. Figure 4.3(a) shows a class map of classified pixels, where color denotes class. It includes two large segments, labeled 1 and 2. One small segment of only two pixels, labeled 3, remains even after majority filtering. In Figure 4.3(b), to remove this small segment, it is reclassified to the most likely class of one of its neighbors, segments 1 and 2, so that it becomes the same class as one of these segments. In the example, the small segment is reclassified to the same class as segment 1, so both segments are denoted in blue, making a new, larger segment.

These steps of pixel classification, majority filtering, and reclassification can be regarded as yielding oversegmented chromosomes since typically it results in more segments than there are chromosomes. In the next section, I discuss how to use these segments to create candidate chromosomes by the process of rejoining.

The resulting segments, the sets $\{S_j\}$, after majority filtering and reclassification, are equivalent to the partition used in Agam and Dinstein's minset representation of the chromosome segmentation problem [16]. The chromosomes can be then formed as minsets of this partition. In the next section, I describe how to choose which set of segments to use to represent each chromosome.

4.2.2 Rejoining of Segments into Candidate Chromosomes

As mentioned, the above steps typically result in oversegmented chromosomes; that is, there is often more than one segment for each chromosome. This is because there are often misclassified segments. In addition, even if pixel classification were performed perfectly, one would need some mechanism to distinguish between which pairs of similarly classified segments represent the ends of one overlapped chromosome and which represent two whole chromosomes within a single cluster. Figure 4.4 illustrates these two possibilities. It shows two clusters of classified segments. Figure 4.4(a) shows a cluster with two segments classified as class 5. In this case, these two segments are two ends of an overlapped chromosome and should be joined together since they are part of the same chromosome. Figure 4.4(b) shows two segments classified as class 6. However, in this case, the two segments are two complete chromosomes and should be recognized as separate.

The rejoining process involves examining all possible pairs of segments as candidate chromosomes and computing the likelihood function $L(\cdot)$ for each of these pairs. The pair that results in the largest likelihood is combined into a single

(a) Overlapped chromosome: two class 5 segments

(b) Two whole chromosomes: two class 6 segments

Figure 4.4: Ambiguity of similarly classified segments

segment, if their likelihood together is greater than the geometric mean of their individual likelihood values. After the pair is rejoined to make a new segment, all possible pairs of the new set of segments are evaluated to find the combination which results in the greatest likelihood. This process is repeated until no more pairs can be found whose combination results in a greater likelihood that the geometric mean of the two original likelihood values in the pair.

The first two pairs are selected as follows:

$$(\tilde{j},\tilde{k}) = \underset{j,k,j\neq k}{\arg\max}\, L\left(S_j^1 \bigcup S_k^1\right) \tag{4.6}$$

The rejoining of two segments is given as

$$S_f^{l+1} = S_{\tilde{j}}^l \bigcup S_{\hat{k}}^l \tag{4.7}$$

where f is an index into a reordered sequence of segments. I repeat this rejoining as long as

(a) Segments after pixel classification and post-processing

(b) Final segments after rejoining

Figure 4.5: Rejoining of segments to make chromosomes

$$\max_i \sqrt{L\left(S_{\tilde{j}}^l\right)L\left(S_{\tilde{k}}^l\right)} < \max_i L\left(S_f^{l+1}\right) \tag{4.8}$$

Since I want to encourage recombining tiny segments which result in a near zero likelihood, the geometric mean is preferred to the arithmetic mean.

When no more pairs can be found suitable for recombination, I have completed the segmentation and classification, and the resulting chromosome segmentation-classification estimates are labeled. Note that abnormal and incorrectly segmented chromosomes will result in a low likelihood and thus can be identified and flagged.

Figure 4.5 shows an example of segment rejoining. Figure 4.5(a) shows the segments that were left after pixel classification, majority filtering, and small segment reclassification. Two segments in the class 6 chromosome have been misclassified as class 14. In this example, the two class 6 segments were joined first, then the lower class 14 segment; then finally, the upper 14 was joined with

the new segment made of the original two 6's and the lower 14. The new large segment was classified as a class 6 chromosome. Joining the class 6 and the class 11 chromosomes did not result in an increase in the likelihood, so rejoining was stopped. The final result is shown in Figure 4.5(b). I have included a flowchart of the algorithm Figure 4.6.

4.3 CONCLUSIONS

In this chapter, I have developed a heuristic for generating a set of candidate chromosomes and a method for obtaining complete chromosomes by rejoining candidate chromosomes using the maximum likelihood hypothesis test. This heuristic uses pixel classification and majority filtering to generate an oversegmented result which is then rejoined using a pairwise maximum likelihood test. The algorithm is able to naturally use multi-spectral information and size with the likelihood function developed in Chapter 3 and account for touching and overlapping chromosomes. Furthermore, it is able to identify abnormal chromosomes. In the next chapter, I present results of applying this algorithm to a set of M-FISH images and compare these results to grayscale segmentation. In addition, I discuss the aberration scoring aspects of the algorithm.

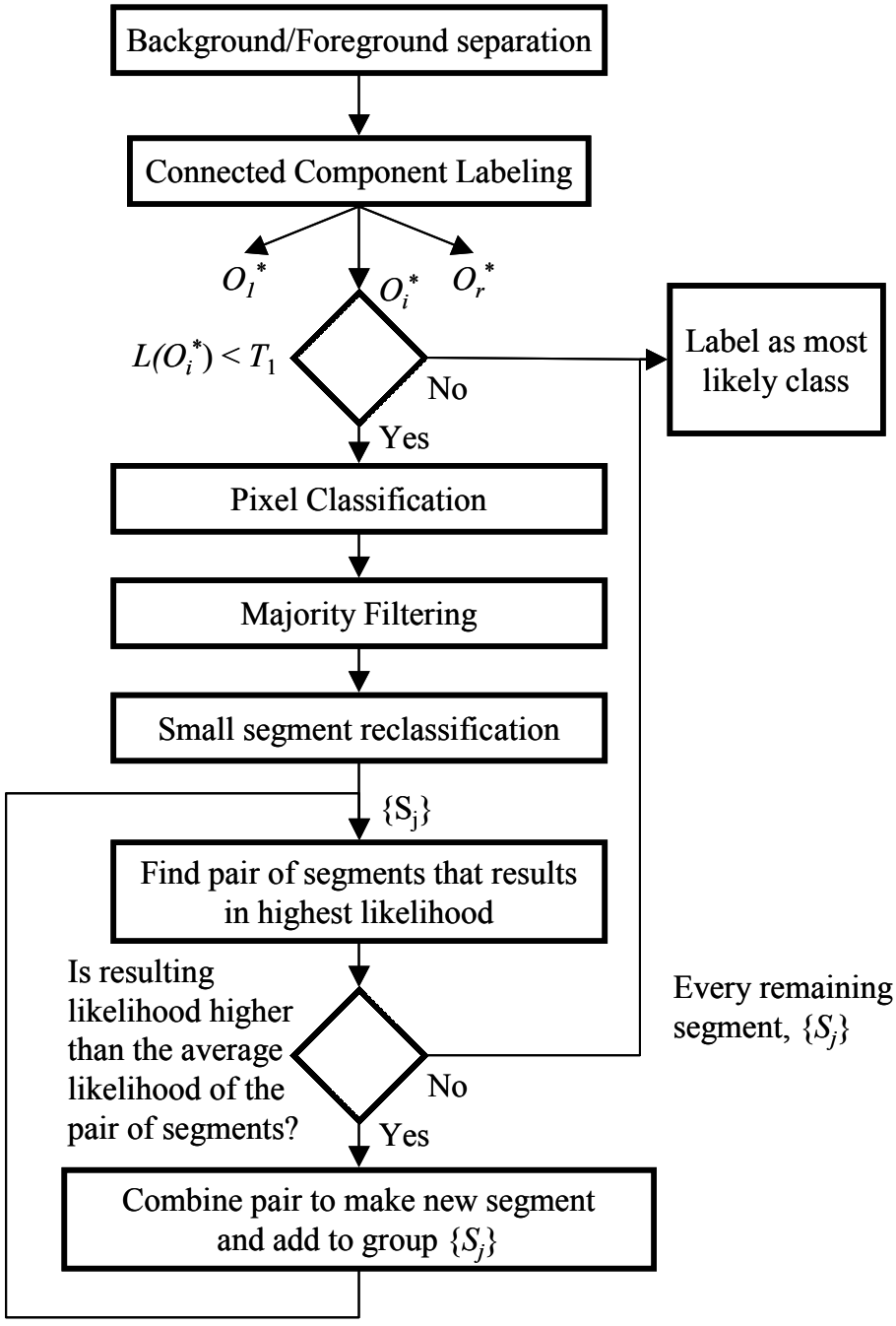


Figure 4.6: Flowchart of proposed segmentation-classification algorithm

Chapter 5: Results

5.1 Introduction

This chapter presents the results of the proposed maximum likelihood joint segmentation-classification algorithm, and examines the strengths and weaknesses of the maximum likelihood joint segmentation-classification algorithm. I compare these results to other current segmentation and classification algorithms. Furthermore, I look at the algorithm's performance as a tool for aberration scoring and error detection.

Section 5.2 describes the set of images on which I tested the maximum likelihood joint segmentation-classification algorithm. Section 5.3 illustrates several examples of the algorithm segmenting images and decomposing clusters of touching and overlapping chromosomes. I compare the segmentation results in Section 5.4 and the classification results in Section 5.5. In Section 5.6, I examine the efficacy of using likelihood to locate different types of abnormalities and to identify segmentation and classification errors. Section 5.7 analyzes the complexity of the algorithm.

5.2 M-FISH Chromosome Image Database

The algorithm was tested on the ADIR M-FISH chromosome image database [50, 51] of 200 multi-spectral images of dimension 517 × 645. Each pixel contains a six element vector of values, five multi-spectral channels plus the grayscale DAPI channel, as discussed in Section 2.6. This database is a

representative set of M-FISH images with a wide variety of image types. It includes images from a variety of dye sets. It includes everything from very simple images with no touches and overlaps between chromosomes to very difficult-to-segment images with a large number of touches and overlaps. It includes crisp, clear images as well as somewhat blurry ones. It includes well spread chromosomes and tightly packed chromosomes. It includes chromosomes at different stages of mitosis. It includes normal male and female karyotypes, sets with simple translocations, and "extreme" cases (labeled karyotype code 'EX') with many abnormal chromosomes.

A nomenclature is used on all the images to easily identify the karyotype of the image and the set to which it belongs. The first character represents the probe set. The next two characters represent the slide number that the image came from. The next two characters are the number of that image on the slide. The final two characters represent the karyotype code. Therefore, if image number 12 from slide 98 (using ASI probes) were from a normal female, its file name would be A9812XX.

The utility of this dataset comes from the fact that it is publicly available and can be used by anyone for comparing M-FISH segmentation results. The dataset also includes an ISCN designation of the karyotype and *a hand-segmented "ground truth" image* for each M-FISH image (marked with a 'K'), so that segmentation results can be easily checked for accuracy.

For this work, I ran the algorithm on every image other than the "extreme" cases. The algorithms were run on each image, and each of the images, together

with the segmentation results compared to the ground truth, or K, images provided with the dataset, were manually analyzed and recorded. Since few segmentations will be pixel-for-pixel identical with the stored K image, this task is somewhat subjective. In all comparisons, I have been rather lenient with both algorithms and accepted any segmentation that varies only slightly from the K image in the database. The results are presented in the following sections.

5.3 EXAMPLES

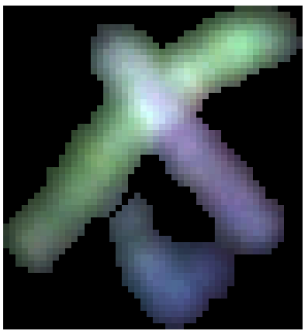
Figure 5.1 shows an example of the maximum likelihood joint segmentation-classification method applied to a single cluster of chromosomes. The cluster includes one touch and one overlap. Figure 5.1(a) shows the original M-FISH image. Figure 5.1(b) illustrates pixel classification using the classifier described in Section 3.3. The effect of the majority filter is shown in Figure 5.1(c). Two small segments were reclassified: the green cluster in the upper left and a single pixel of red in the class 22 chromosome. Figure 5.1(d) shows the segments after reclassification and the likelihood values of their most likely class. Figure 5.1(e) shows the final segmentation. The ends of the class 15 chromosome have been rejoined, as have the two segments of the class 22 chromosomes. Notice that the likelihood of the class 12 chromosome is still low since it covers part of the class 15 chromosome.

Figure 5.2 shows an example of a simple image (A0105XY) segmented with the maximum likelihood joint segmentation-classification method. Again I have shown pixel classification, majority filtering, and the final segmentation after rejoining. In this image, there were no segments small enough for

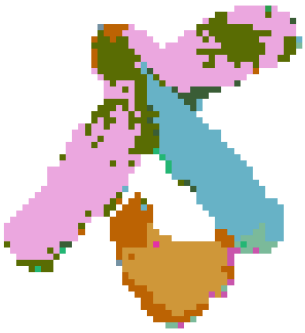
reclassification. Small segments, in fact, are rather rare given the large size of the majority filter, but they are still possible, as witnessed in Figure 5.1. So, I include small segment reclassification in the algorithm for completeness. In this example, two touches and the overlap were decomposed correctly.

Another example is given in Figure 5.3. This image (V190542) has no overlaps, but is notable because its chromosomes are tightly packed and it contains several multi-chromsome clusters, which can be difficult for some algorithms to segment. In addition to its chromosomes being close together, this image is abnormal in its number of chromosomes, having four extra chromosomes. In this example, all touches were correctly decomposed, although two chromosomes were misclassified. Two of the chromosomes classified as 8's should have been 7's. Type 7 and type 8 chromosomes are similar in size, as shown in Table 2.1. This, together with errors in pixel classification, led to them being misclassified.

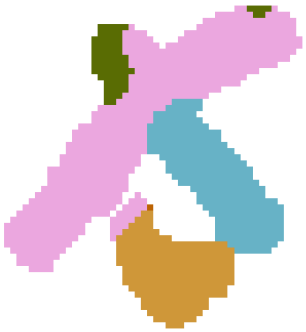
In Figure 5.4 and Figure 5.5, the strengths of each method are contrasted. Figure 5.4 shows an example where the maximum likelihood M-FISH method succeeds where grayscale methods fail. In this example, two chromosomes touch closely and are in line with each other. Using grayscale and geometric information alone, this appears to be a single chromosome, where the M-FISH multi-spectral data makes the touch clear. Figure 5.5, on the other hand, shows an example where multi-spectral methods fail. In this example, two chromosomes of the same class are overlapping. The geometric information succeeds here, but the multi-spectral information cannot distinguish between the two chromosomes.



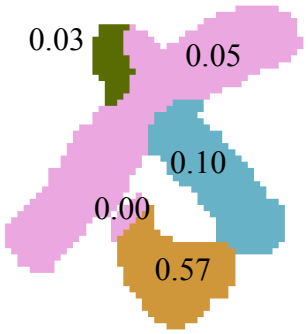
(a) M-FISH cluster



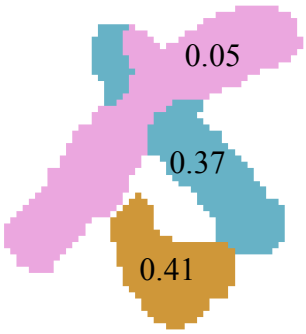
(b) Pixel classification



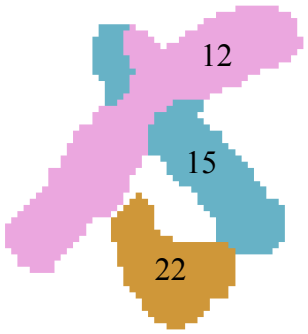
(c) Majority filtering



(d) Small segment reclassification with segment likelihood values

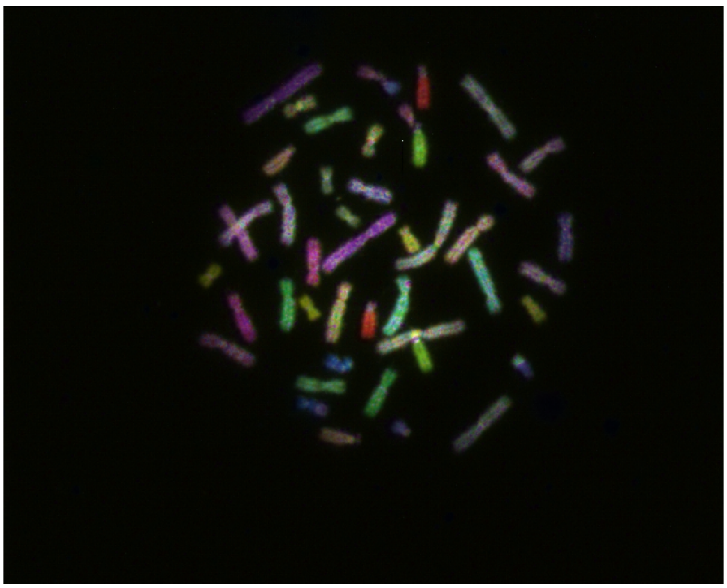


(e) Final likelihood values



(f) Final classification

Figure 5.1: Example of cluster decomposition



(a) M-FISH image

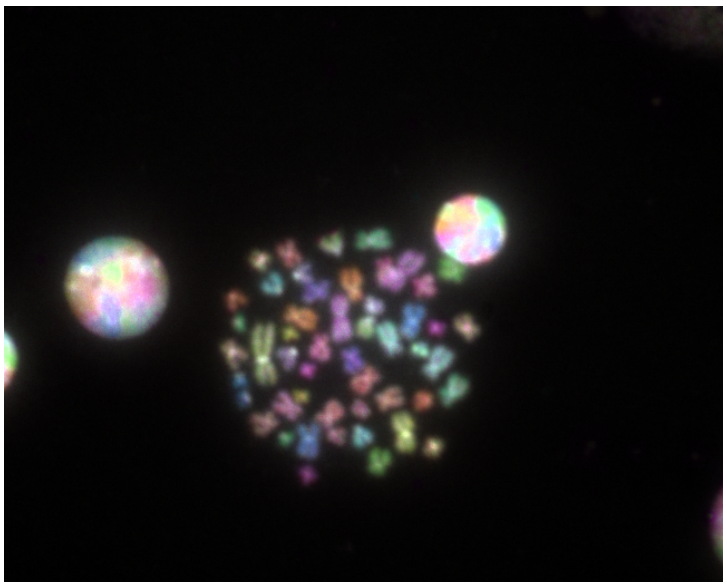
(b) Pixel classification

Figure 5.2: Example of M-FISH image segmentation

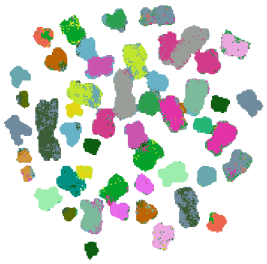
(c) Majority filter

(d) Final segmentation-classification

Figure 5.2: Example of M-FISH image segmentation

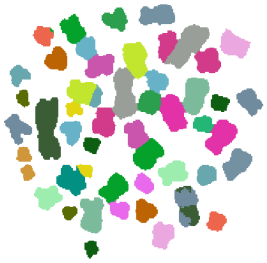


(a) M-FISH image



(b) Pixel classification

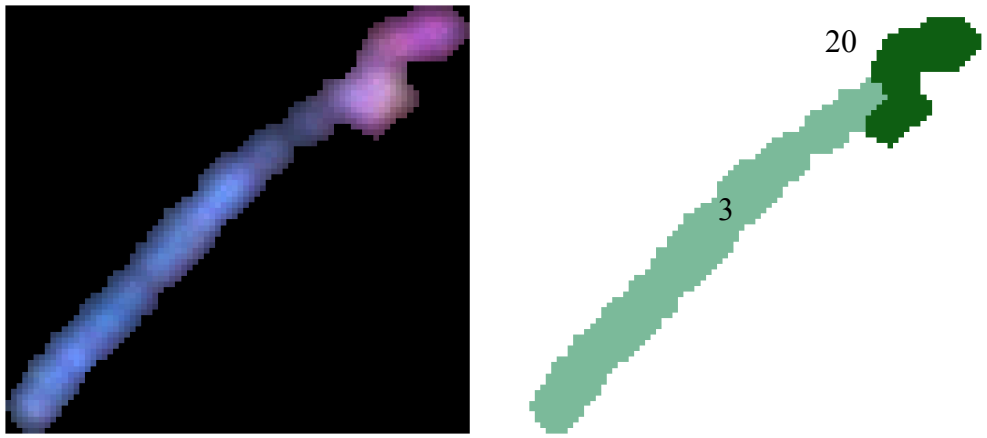
Figure 5.3: Another example of M-FISH image segmentation



(c) Majority filter

(d) Final segmentation-classification

Figure 5.3: Another example of M-FISH image segmentation



(a) M-FISH image

(b) Maximum liklehood segmentation-classification

Figure 5.4: Multi-spectral methods work, but grayscale methods do not.

(a) M-FISH image

(b) Grayscale segmentation

Figure 5.5: Grayscale methods work, but multi-spectral methods do not.

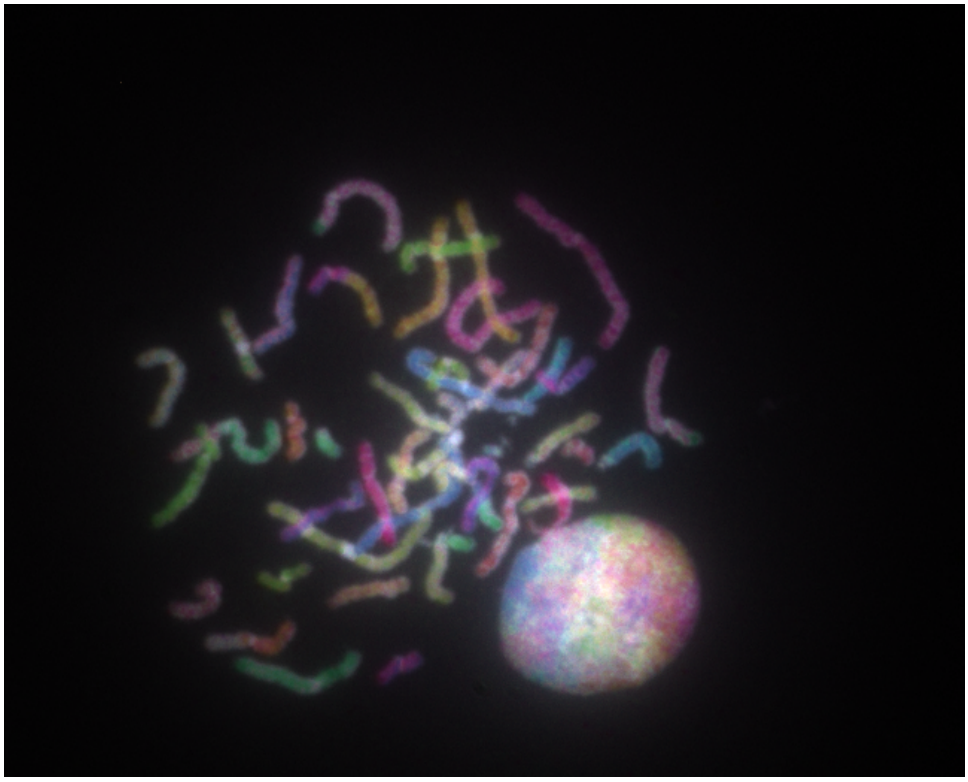


Figure 5.6: M-FISH image that is difficult to segment because of the many overlapping and tightly packed chromosomes.

While most of the examples in this section were straightforward and successfully decomposed, the ADIR M-FISH dataset includes a wide variety of images, including a number of difficult real-world images such as Figure 5.6. In the next section, I discuss the algorithm performance on the whole database and examine several types of clusters which the algorithm does not segment correctly.

5.4 SEGMENTATION

I compare my segmentation results against the Cytovision chromosome segmentation software [52], a popular, commercially-available package of chromosome imaging software that performs grayscale image segmentation. I

applied the software to the DAPI channel of each image in the ADIR chromosome image dataset. The Cytovision software is semi-automatic. It first decomposes what it believes are certain touches, then marks what it believes to be more possible clusters (see Figure 5.7), and the user manually selects the touches and overlaps. It then attempts to decompose the touches and overlaps that the user selects. For comparison, I have manually selected all touches and overlaps, to see what percentage are correctly decomposed. If a touch or overlap is segmented incorrectly, I left it unsegmented, rather than let it be segmented incorrectly. For this reason, the Cytovision grayscale software has many more clusters unsegmented than segmented incorrectly. If a cluster contained both touches and overlaps, I manually selected the order of decomposition that resulted in the best segmentation.

The maximum likelihood method runs completely automatically. It attempts to recognize all clusters and decompose them, treating touches no differently than overlaps, since they are both decomposed in the same way in the algorithm.

Since I have not concerned myself with background/foreground separation in this work, I will assume ideal separation. For the Cytovision grayscale software, I manually selected the threshold that resulted in the best segmentation accuracy. Cell nuclei and debris were removed manually. For the maximum likelihood algorithm, I used the background/foreground separation included in the hand-segmented K file of the M-FISH image dataset. If these two different

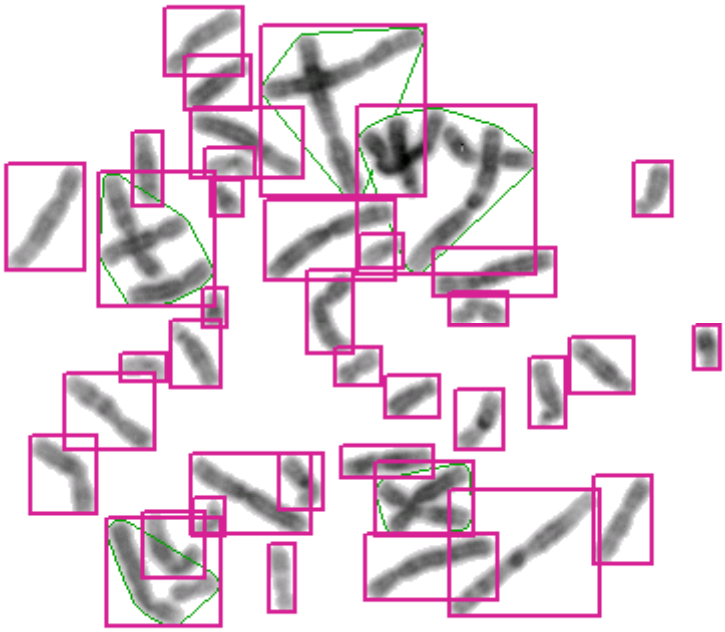


Figure 5.7: Cytovision interface. Detected clusters marked in green outlines.

methods resulted in different clusters, those clusters were discarded, and only matching clusters were compared.

Since the Cytovison grayscale software requires human assistance, a comparison between it and the maximum likelihood method might not be completely fair. For instance, very few clusters are oversegmented or missegmented in the grayscale software since the user only selects decompositions that are performed correctly. The only oversegmented and incorrectly segmented clusters results from the software automatically segmenting clusters it believes to be obvious. To account for some of this, I have calculated

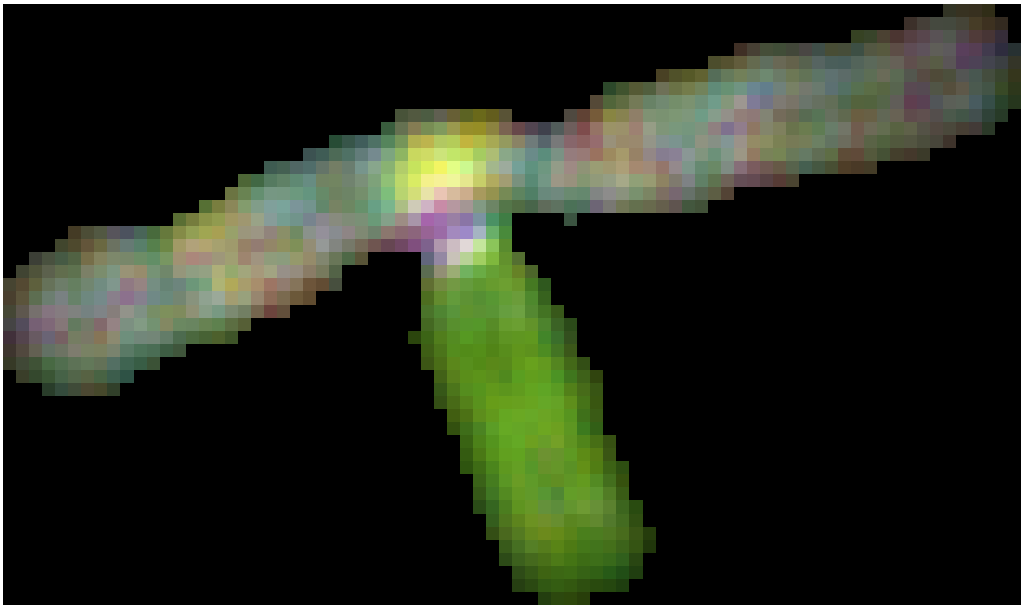


Figure 5.8: “Hard” touch. Only the tip of a chromosome is overlapped, so unlike the typical overlap case, both ends of the chromosome are not visible.

the percentage of clusters and single chromosomes that both methods have identified as clusters.

The segmentation results are shown in Table 5.1. The maximum likelihood method correctly decomposed a much higher percentage of touches compared to the grayscale segmentation. The grayscale segmentation particularly has a difficult time with “hard” touches, or partial overlaps (see Figure 5.8), and with large clusters of many tightly packed chromosomes (see Figure 5.3). Neither does very well with overlaps, although the grayscale method seems more reliable that the maximum likelihood classification-segmentation. This is partly because the probabilistic modeling resists overlaps since it is uncertain about the part being overlapped. It often mistakes overlaps for a touch.

Most chromosomes incorrectly segmented by the maximum likelihood joint segmentation-classification algorithm fall into one of five classes:

1. The most obvious class is the example shown in Figure 5.5. If two chromosomes of the same class touch or overlap, there is no way to determine their boundary with multi-spectral information alone. Grayscale or geometric information must be used in this case.
2. In certain instances a chromosome or cluster of chromosomes was incorrectly segmented because of poor background/foreground segmentation. As mentioned, I did not perform my own background/foreground segmentation, but instead use the background/foreground segmentation that was contained in the K files of the M-FISH image dataset. While the segmentation in the K files always contains the chromosomes, it does not necessarily guarantee that the masks will exactly match the border of the chromosomes. The masks can be larger than the chromosomes, sometimes twice as large as the chromosomes they contain. Because segmentation likelihood is a function of size, incorrect size information derived from the background/foreground segmentation can lead to erroneous results. Figure 5.9 shows the border of the K files background/foreground segmentation in white. The border hugs tightly to the edge of the chromosomes in most instances, but the bottom half of the green

Figure 5.9: Background/foreground inaccuracies in K files

chromosome was manually enlarged by the dataset's creators to include the telemere.

3. Translocations within a cluster of touching or overlapping chromosomes inevitably lead to incorrectly segmented chromosomes. Because both translocations as a whole, and its individual segments, often yield a low likelihood, the algorithm will not be able to find any segmentation of high probability for it. This can often lead to a chain reaction of other chromosomes in the cluster being incorrectly segmented because the low likelihood segments in the translocation may now pair with any other segment in the cluster.

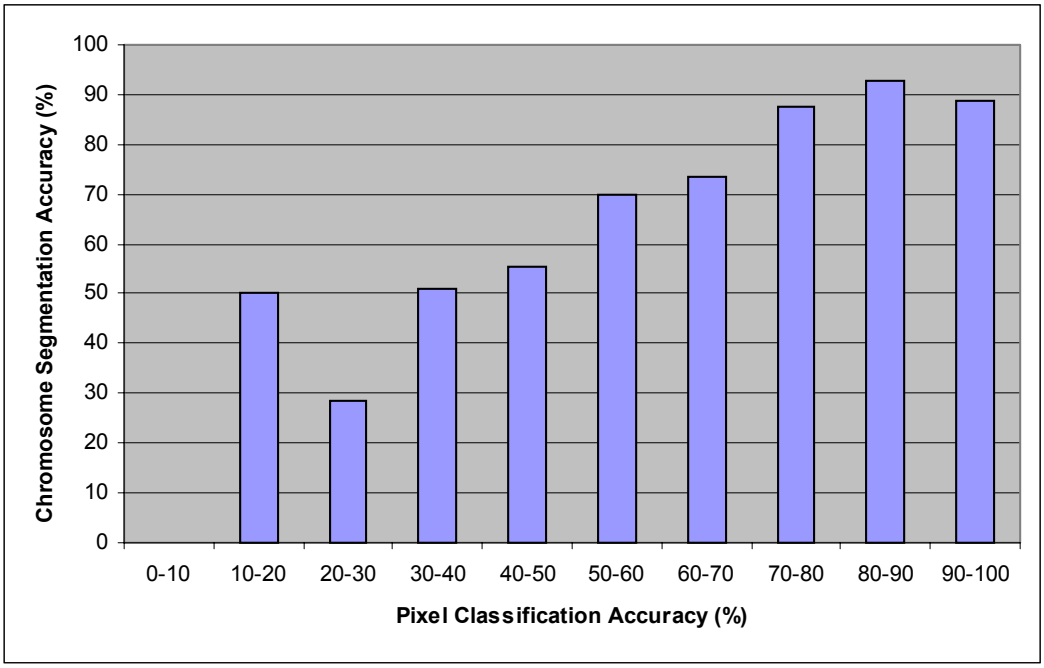
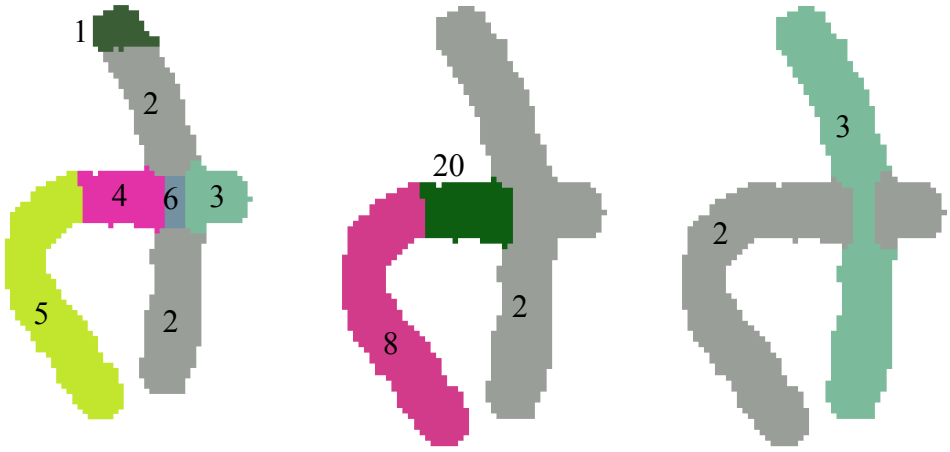


Figure 5.10: Impact of pixel classification on segmentation

4. Of course, segmentation-classification accuracy is inherently dependent on pixel classification accuracy. Pixel classification rates vary widely throughout the M-FISH image dataset. Average pixel classification accuracy for my classifier was 68% with a standard deviation of 17.5%. Accuracies above 90% were not uncommon for some images, and some images only had pixel classification accuracies of 20-30%, or even less in a few rare cases. Figure 5.10 shows a chart of segmentation-classification accuracy versus pixel classification accuracy. The "10-20" bar in this graph is statistically insignificant because there are only a few images in this group.

5. Finally, a common cause of errors in segmentation-classification was the "greedy" approach to the algorithm. The algorithm does not guarantee an optimal combination of segments in the sense of likelihood. Instead it is a greedy algorithm, in that it only rejoins the pair that results in the highest likelihood for a single combination. It is possible that rejoining two segments with a lower rejoined likelihood may lead to a series of rejoinings that have an overall higher likelihood. In Figure 5.11, the pair of segments that results in the highest likelihood is segments 2 and 3. Segment 3 produces a higher likelihood than 1 and 6 since it is larger and makes the chromosome closer to its expected size for a class 3 chromosome. Also it includes some of the class 3 chromosome, so its multi-spectral information might also match somewhat. However, with that approach, while one segment of high probability is found, there is no combination for segments 4 and 5 that will result in a high probability, so the overall average likelihood is low.

In addition to decomposition accuracy, another important factor to consider is an algorithm's accuracy in detecting clusters, so that it will know which objects it should keep as single chromosomes and which ones it should attempt to decompose into parts. One can see in Table 5.2 that the probabilistic model recognizes a much higher percentage of clusters, but as a result also recognizes more single chromosomes as clusters. However, even though the



(a) Numbered segments (not classes)

(b) Average likelihood of greedy segmentation: 0.08

(c) Average likelihood of optimal segmentation: 0.28

Figure 5.11: Greedy vs. optimal

probabilistic model recognizes 6% of singles as possibly clusters, less than 1% of them are actually oversegmented when they should not have been (Table 5.1), because very rarely will any segments within a single chromosome result in a higher probability than the chromosome as a whole. In this test, I use a threshold of 0.1 likelihood for determining whether a connected component or not; if a segment has less than a 0.1 likelihood of belonging to its most likely class, I assume that that object is a cluster or an abnormality.

5.5 CLASSIFICATION

Classification accuracy was also run on the entire ADIR M-FISH database of 200 images. I examined all chromosomes correctly segmented by the maximum likelihood method, both in clusters of touching and overlapping

Table 5.2: Objects recognized as clusters. ML method recognizes more clusters, but also incorrectly recognizes more single chromosomes as clusters. However, the ML method only actually oversegments 0.8% of single chromosomes compared to 0.2% for the Cytovision method.

Recognized as Clusters	***Count***	***ML Method***	***Cytovision***
Clusters	496	95%	69%
Singles	3102	6%	0.4%

chromosomes and by themselves. Incorrectly segmented chromosomes, translocations, and other abnormal chromosomes were not considered. Then the M-FISH database K files were used to determine whether the classification was correct or not.

For comparison, I show the misclassification rate for chromosome classification using pixel classification alone. The method I used for classifying chromosomes based solely on pixel classification was to classify each chromosome to the class that occurs most often in the classified pixels within that chromosome. Table 5.3 shows that this chromosome classification, based only on pixel classification, has almost twice the misclassification rate than the proposed likelihood function $L(\cdot)$ and the joint segmentation-classification algorithm.

5.6 CHROMOSOME FLAGGING

One advantage of the maximum likelihood joint segmentation-classification algorithm is that the final result of each segmentation and classification is a likelihood value for each chromosome segment, and the likelihood value is a measure of the certainty of the classification of that segment.

Table 5.3: Chromosomes classification accuracy. Using the proposed likelihood function for classification reduces misclassifications by nearly 50% compared to classification using only multi-spectral data.

	Joint Segmentation-Classification	***Only pixel classification***
Misclassified	8.1%	15.0%

This measure of certainty is useful because it allows segments of low likelihood to be flagged and presented to a human user as segments that might require more manual inspection.

There are four possibilities that would result in a segment of low probability:

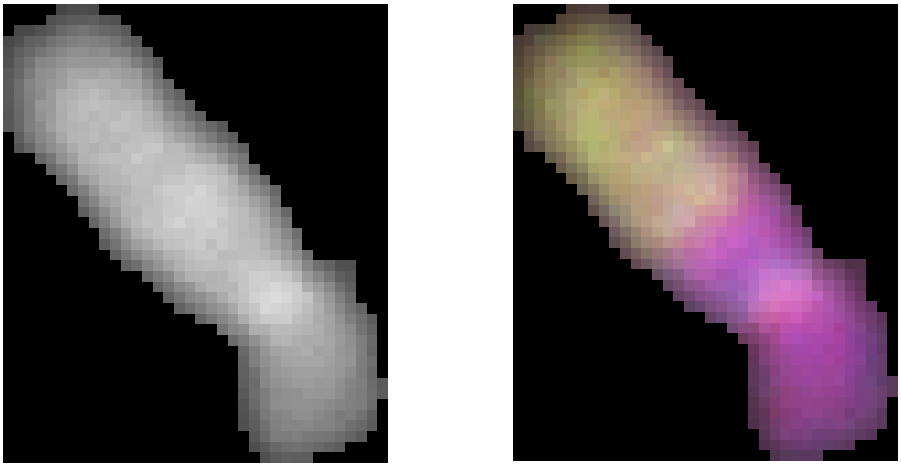
1. The segment is a translocation, broken chromosome, or some other abnormal chromosome. In this case, there is no "correct" segmentation since the chromosome as a whole will not match the likelihood function's idea of a chromosome, and the sections of a translocation will be too small to receive a high likelihood measure.
2. The segment is incorrectly segmented. If the maximum likelihood method errs and cannot find the correct segment, the resulting segments often will have low likelihood.
3. The segment is misclassified. Even if segmented correctly, noise, weak dyes, or other factors could cause the segment to be misclassified. In this case, the likelihood function will also be low since likelihood function $L_1(\cdot)$, which measures pixel classification certainty, will be low.

4. It is possible that the segment is segmented and classified correctly, but that the segment still has low likelihood. This may also be due to noise, weak dyes, image distortion, misregistration of spectral images, etc.

In the first three of these four cases, it is useful to flag these segments and present them to the user so that the user can either fix the segmentation-classification error or further inspect the abnormal chromosome. In practice, all karyotypes are reviewed manually, but such flagging, or ranking of segments by likelihood, would certainly save time since the user is automatically directed to the questionable segments, rather than having to examine every segment for correctness without any prior knowledge. In the following sections, I examine these four possibilities.

5.6.1 Aberration Scoring

Possibly the most important aspect of karyotyping is anomaly detection. Extra chromosomes, missing chromosomes, and translocations are indicators of radiation damage, cancer, and a wide variety of genetic disorders. Because of the multi-spectral information in M-FISH images, many types of anomalies, such as translocations, that were not detectable in grayscale images are readily apparent [45]. Figure 5.12 shows an example of a t(20:5) (the standard designation [7] for translocation between a type 20 chromosome and a type 5 chromosome). Even an untrained observer can easily see the translocation in the M-FISH version because of the significant difference in color of the two sections. However the grayscale version appears as a normal chromosome, at least to an untrained observer.



a) Grayscale b) M-FISH

Figure 5.12: t(20;5) translocation. An exchange of material between a type 20 and type 5 chromosome.

The proposed probabilistic model for segmentation-classification also aids in locating these translocations, since the translocations and their segments result in low probability. If a translocation has two segments, both of these segments individually will be too small to have a reasonable probability for that segment's class, and both segments together will often be too large for either segments class. So these abnormalities will quickly be located and their severity measured. Broken chromosomes are also easily identified since their broken segments will be too small to result in a reasonable probability. Thus for translocations, broken chromosomes, and other abnormalities, likelihood can be used as a criterion for identifying abnormalities. If segments of low probability are found, the image and its karyotype can then be marked as abnormal for later examination by a human expert. Of course, locating abnormalities solely based on likelihood will not help flag an abnormal number of chromosomes, unless chromosome count is

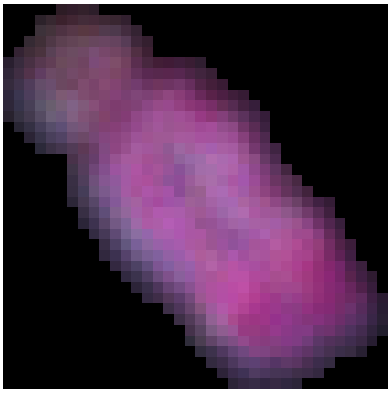


Figure 5.13: Small translocation; t(7;8)

somehow incorporated into the likelihood function. However, it is a simple matter to flag a class if it has more (or less) than its normal number of chromosomes assigned to it, so I will ignore the case of abnormal number for these tests.

It is important to note here than some translocations are very difficult to detect, even for an expert, since they are very small. It is not always clear whether a tiny change in color at the end of a chromosome is due to noise or staining, or is an actual translocation. Sometimes it requires several images from the same patient to verify that a chromosome actually contains a translocation. Figure 5.13 shows a t(7:8) which is quite similar to a normal class 8 chromosome. It is much less noticeable than the translocation in Figure 5.12 because it only has a small section of class 7 chromosome, and it is similar in size to a normal class 8 chromosome. Many translocations are even less noticeable.

The proposed likelihood function also can have difficulty in detecting these small translocations because they change the size of the chromosome by

Table 5.4: Abnormality detection characteristics on V29 image set in the ADIR M-FISH dataset. On average the likelihood value for a translocation is significantly lower than the value for normal chromosomes.

	Normal Chromosomes	***Translocations***	***Fragments***
Likelihood average	0.44	0.12	0.02
Likelihood standard deviation	0.24	0.10	0.02
< 0.1 likelihood	4.9%	50%	100%
< 0.3 likelihood	34%	96%	100%

only a small amount, and because there is still high confidence of the class throughout most of the chromosome. The proposed likelihood function is quite reliable, though, in detecting larger translocations and smaller than normal chromosomes that might result from a break.

Table 5.4 shows the results of running the algorithm on image set V29 from the ADIR M-FISH dataset. This dataset has 15 images with 5 translocations in each, as well as some short chromosome fragments. For this test, I compare the likelihood values of the normal, correctly segmented chromosomes to the likelihood values of the abnormal chromosome material, the translocations and the fragmented chromosomes. One can see how effective the likelihood value is as a feature for distinguishing between normal and abnormal chromosomes. The average likelihood values of the translocations and the partial chromosomes are much lower than the average probabilities of the whole chromosomes. This table also shows what percentage of normal and abnormal chromosomes were flagged

Table 5.5: Likelihood function < 0.1. The proposed likelihood function is much more likely to flag abnormals and errors in segmentation and classification than normal, correctly identified chromosomes.

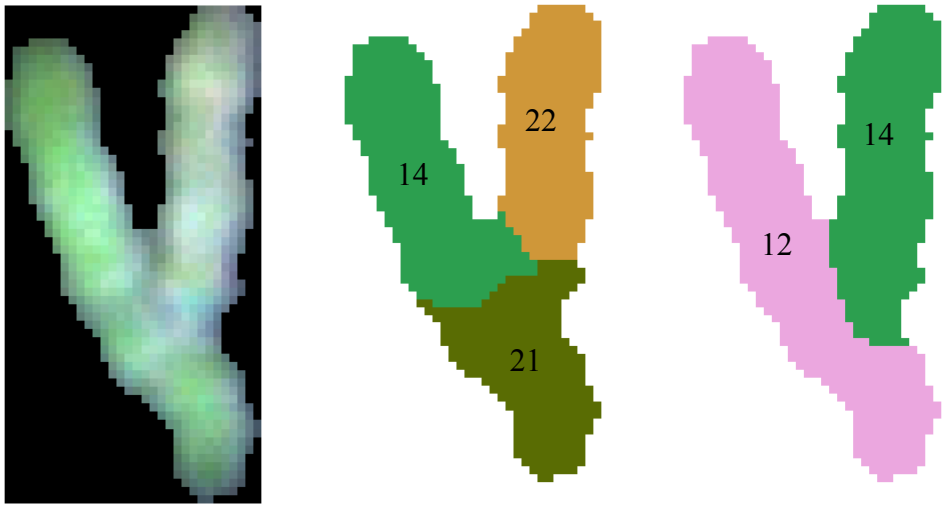
	Count	***Flagged***
Abnormals	114	49.1%
Incorrect Segmentation	409	52.6%
Incorrect Misclassification	315	48.6%
Correct Segments	3866	6.4%

with a likelihood threshold of 0.1. This number is arbitrary and used only the purpose of illustrating the disparity between the percentage of normal chromosomes flagged and the percentage of abnormal chromosomes flagged. In fact, the threshold for abnormality detection should probably be higher than 0.1 since the average likelihood for a translocation is 0.12, so I have also included a higher threshold of 0.3. This higher threshold flags almost all the abnormal chromosomes, but also catches more normal chromosomes.

Table 5.5 shows the results of abnormality detection on the entire database. With a likelihood threshold of 0.1, 49.1% of all the abnormal chromosomes in the database are flagged.

5.6.2 Incorrect Segments

While it is hoped that all chromosomes are segmented perfectly, this is not always the case. However, if it were not possible for an algorithm to find the correct segmentation, one would hope that it would be able to point out the questionable segments for human inspection. Because of the likelihood function,



(a) M-FISH cluster

(b) Incorrect segmentation and classification

(c) Correct segmentation and classification

Figure 5.14: Single flagged segment can correct a whole cluster. Cluster is incorrectly segmented and classified. However, flagging only one segment can direct a user to correct the whole cluster.

this is possible with the proposed maximum likelihood algorithm. If the maximum likelihood function cannot find a likely segmentation possibility, it results in a low likelihood and the segments involved are easily located.

Table 5.5 shows the percentage of incorrect segments that are flagged. Again I use 0.1 as my threshold for flagging segments. While the table shows that only 52.6% of incorrect segments are flagged, the algorithm may be more effective than this number might indicate, since many of the cases are part of an incorrectly segmented, multiple chromosome cluster. So while only one segment in that cluster might be flagged, this is effectively the same as flagging the cluster, since fixing that segment will likely fix other segments in that cluster that might

not have been flagged. Figure 5.14 shows an example of this. Figure 5.14(b) shows a cluster that has been incorrectly segmented and classified. If segment 14 or 21 were the only segment flagged, they would result in the whole cluster being corrected, because one cannot be fixed without the other since they are both part of the same chromosome. And since the correct chromosome, 12, would then be correctly segmented, the remaining part of the cluster, the class 14 chromosome, would also be corrected.

5.6.3 Misclassifications

Finally I looked at misclassifications. Misclassifications also generally result in a low likelihood. This is because misclassifications often result from uncertainty caused by noise or weak labeling. Very rarely is a chromosome misclassified with a high probability. Table 5.5 shows that likelihood is, in fact, a good indicator of misclassified chromosomes, with 48.6% of misclassified chromosomes having a likelihood of less than 0.1.

As in the case of abnormal chromosomes, the algorithm would certainly flag a higher percentage of chromosomes if I raised this likelihood threshold, but it would also flag a higher percentage of correct chromosomes. This might, in some instances, be desirable if one were willing to sort through more correct segments in order to catch more incorrect or abnormal ones. However, the arbitrary threshold of 0.1 has been used in these examples just to illustrate the disparity of flagging between abnormal chromosomes and incorrect segments compared to correct segments, which are presented in the next section.

5.6.4 Correct Segments

For comparison, I also include the rates for correct segments. One can see in Table 5.5 that only 6.4% of correctly segmented and classified chromosomes are flagged. This means that correct segments only constitute 38% of the flagged chromosomes, even though there are almost 5 times as many correct segments than abnormal chromosomes, incorrectly segmented chromosomes, and misclassified chromosomes put together.

5.7 COMPLEXITY

The lengthiest part of the algorithm is the pixel classification (Section 3.3). It requires evaluation of a multi-dimensional Gaussian for each class and each pixel to calculate the probability of each pixel belonging to each class. Since there are 5 dyes, this requires a 5 element mean vector be multiplied by a 5×5 covariance matrix multiplied by another 5 element mean vector, followed by a single exponential. This is a total of 30 multiples and 24 adds for each pixel and class. Since there are 24 classes, and the standard image size used is 517×645, this results in almost 240 million multiplies and adds for each image. This also makes it the most memory intensive part of the algorithm, since it must store a probability for each pixel and each class.

The majority filter (Section 4.2.1) involves counting the number of occurrences of each class within a filter. Since the filter used in this work is 17×17, with a naïve approach 289 increments of the proper class counter would be necessary for each pixel. However, with a raster-scanned image, it is only necessary to adjust the previous class counts by incrementing the counters with

the front edge of the filter and decrementing in with the trailing edge of the filter. With this implementation, for a 517×645 image, there would be 5.6 million increments and 5.6 million decrements per image for the majority filter.

The likelihood calculation (Section 3.3) for any segment is relatively simple. The first likelihood function only involves averaging the probabilities over a segment. Thus it only involves one add for each pixel in the segment, and a single divide. The second likelihood is simply the evaluation of a single-dimension Gaussian. Estimating the area of overlap is of $O(x^2)$ complexity as a function of the number of boundary pixels, but rarely involves a number of boundary pixels between chromosomes greater than 50, so it also can be computed quickly.

The merging algorithm (Section 4.2.2) is of $O(x^2)$ complexity since it compares every pairwise combination of segments. In general, most clusters can be completely merged within a few seconds, but because of the algorithm's $O(x^2)$ complexity, clusters with many segments can take significantly longer.

The major memory components necessary in the algorithm are the M-FISH image (6 images) and the pixel-classification probabilities (24 images). In addition, a few other images must be stored in memory such as the classified pixel map, the connected component labeling, the majority filter output, and the image of the output segments.

The code for the algorithm presented in this dissertation is available at `http://signal.ece.utexas.edu/~wade/mfish`. For a typical 517×645 image, this code takes around 2.5 minutes on a 167 MHz Sun

workstation. Most of that time is used for pixel classification and its probability calculations. In addition, a portion of that time is used for calculating side information not strictly necessary for the operation of the algorithm.

5.8 CONCLUSIONS

This chapter examined the performance of the proposed maximum likelihood joint segmentation-classification algorithm on multi-spectral M-FISH images. It showed a 33% improvement in decomposing touching chromosomes over grayscale chromosome segmentation methods that only used information from the DAPI M-FISH dye. The proposed maximum likelihood method especially excelled with "hard" touches and difficult, tightly packed clusters with many chromosomes. It also shows improved performance over past chromosome classification methods which only use pixel classification information, having only about half the number of misclassifications. Finally likelihood values were shown to be a reliable indicator of incorrect segments and abnormal chromosomes, since, with a likelihood threshold of 0.1, abnormal chromosomes and incorrect segments are around 8 times as likely to be flagged as normal segments.

Chapter 6: Conclusions

A method to segment and classify chromosomes in M-FISH images based on multi-spectral information is introduced in this dissertation. It uses pixel classification and a probabilistic model of chromosome features to select from among a set of segmentation possibilities. Since the model is a function of both segmentation and classification, both can be achieved simultaneously. The method is able to decompose both overlaps and clusters composed of more than two chromosomes. Furthermore, since this method, in general, is not specific to the shape or characteristics of chromosomes, it could possibly be used on other multi-spectral segmentation problems where different objects in an image have different spectral signatures.

Chapter 2 introduces chromosome imaging and the concept of karyotyping. I discuss various characteristics of chromosome images and features of chromosomes that are used for karytyping and classification. Finally M-FISH imaging is introduced and it is shown that the new multi-spectral chromosome features available in this imaging modality may be useful for better segmentation and classification.

In Chapter 3, I develop the theory and notation for a unified segmentation-classification system. I write a likelihood function that measures the quality of both segmentation and classification. By maximizing this likelihood function, one can accomplish both segmentation and classification in one step. I also develop a likelihood function appropriate for the specific application of M-FISH

chromosomes segmentation and classification. The likelihood function uses the features of multi-spectral information, chromosomes size, and estimated overlapped area. It is proposed that the framework could work as well for other applications, given a likelihood function appropriate to that application.

The framework in Chapter 3 only provides a way of measuring and comparing different segmentation possibilities. In Chapter 4, I introduce a method to generate a set of segmentation possibilities for evaluation by the likelihood function. This method involves pixel classification and several steps of post-processing to acquire an oversegmented image. Clusters in the image are then detected, and the segments in those clusters are combined in the way that most improves the cluster's likelihood. In this way, clusters of both overlapping and touching chromosomes can be decomposed.

The performance of the algorithm is examined in Chapter 5 by applying the algorithm to the ADIR M-FISH image database. The proposed maximum likelihood method is shown to give a significant improvement over past grayscale segmentation techniques in decomposing clusters of touching chromosomes. In addition, the proposed algorithm was able to detect clusters of chromosomes more reliably, without significant sacrifice in terms of allowing oversegmented chromosomes. In classification performance, the algorithm was shown to give only about half the number of misclassifications as classification by classified pixel counting alone. Finally, the likelihood function was shown to be a reliable indicator of abnormal segments such as translocations, incorrect segmentation, and misclassified segments.

Future research in M-FISH chromosome imaging could involve developing a more intricate likelihood function for M-FISH imaging. While the features used in this work are simple to measure and calculate and have shown themselves to be reliable indicators of segmentation-classification quality, it is likely that a more detailed likelihood function with more features might lead to an even better measure. As suggested in Chapter 3, the likelihood function could include other components such as likelihood functions based on size or chromosome number. Another possibility for modifying the likelihood function could involve using prior information from previous images for locating abnormalities in the current image. For instance, if the previous image had a translocation, then the current image might be likely to have one as well, involving the same classes.

Further research in pixel classification methods would also be productive area of research. Given the algorithm's dependence on pixel classification accuracy (Figure 5.10), it is clear that improvements in pixel classification accuracy will, in turn, result in improvements in chromosomes segmentation-classification accuracy. Methods of background/foreground segmentation could also be further investigated, since they also the impact they have on segmentation-classification accuracy.

The suboptimality of the proposed algorithm due to its "greedy" nature has already been illustrated in Figure 5.11. Another method of combining segments will be necessary to achieve optimal segmentation in the sense of the likelihood function. While the difficulties with cutlines and similar techniques for

has already been documented (Figure 4.2), another method altogether of developing segmentation possibilities, possibly not even based on classified pixels and segment rejoining, might also assist in realizing this optimality.

Finally, it will be important to find the best way to combine multi-spectral based segmentation methods with grayscale methods. Since M-FISH techniques generally include a DAPI channel (see Section 2.6), a grayscale representation of the chromosome image is available as well as the 5-channel multi-spectral representation. As it has been shown in Figure 5.4 and Figure 5.5, there are clusters that are better segmented with multi-spectral information and clusters that are better segmented with grayscale and geometric information. Clearly both types of information are necessary for the most accurate segmentation. For instance, valley-searching [2] techniques typically produce a more accurate boundary than pixel classification based segmentation techniques. However, valley searching techniques are not necessarily as useful for detecting clusters or for choosing the more likelihood of two completely separate segmentation possibilities. For a complete M-FISH chromosome segmentation-classification algorithm to be developed, it will need to be able to cleanly integrate all these different types of information.

Appendix: M-FISH Labeling Charts

Courtesy Advanced Digital Imaging Research, LLC, League City, Texas 77573

ASI M-FISH (SKY) Kit

Chromosome Class	Spectrum Green	Spectrum Orange	Texas Red	Cy5	Cy5.5
1	x		x	x	
2					x
3	x	x		x	x
4	x			x	
5	x	x	x		x
6	x		x	x	x
7			x	x	
8	x				
9	x	x			x
10				x	x
11	x	x		x	
12			x		x
13	x	x			
14			x		
15		x	x	x	
16	x		x		
17				x	
18	x	x	x		
19		x		x	
20		x			
21	x				x
22		x	x	x	x
X		x			x
Y	x			x	x

PSI M-FISH Kit

Chrom.	Deac	FITC	532	568	Cy5
1			x		
2				x	
3	x				
4		x		x	
5			x		x
6		x			
7			x	x	
8				x	x
9					x
10	x		x		x
11	x			x	
12		x	x		
13	x	x			
14		x	x	x	
15	x		x	x	
16		x			x
17		x		x	x
18			x	x	x
19		x	x		x
20	x			x	x
21	x	x	x		
22	x	x		x	
X	x				x
Y	x		x		

Vysis M-FISH kit

Chrom.	Spectrum Aqua	Spectrum Green	Spectrum Gold	Spectrum Red	Far Red
1			x		
2				x	
3	x				
4		x		x	x
5			x		x
6		x			
7					x
8				x	x
9			x	x	
10	x		x		
11	x			x	
12		x	x		
13	x	x			
14		x	x	x	
15	x		x	x	
16		x			x
17		x		x	
18			x	x	x
19		x	x		x
20	x			x	x
21	x	x	x		
22	x	x		x	
X	x				x
Y	x		x		x

Bibliography

[1] M. R. Speicher, S. G. Ballard, and D. C. Ward, "Karyotyping Human Chromosomes by Combinatorial Multi-fluor FISH," *Nature Genetics*, vol. 12, pp. 368-375, 1996.

[2] A. M. Vossepoel, *Analysis of Image Segmentation for Automated Chromosome Identification*, University of Leiden, Leiden, Netherlands, Doctoral Dissertation, 1987.

[3] Human Chromosome Study Group, "A Proposed Standard of Nomenclature of Human Mitotic Chromosomes," *Cerebral Palsy Bulletin*, vol. 2, no. 3, 1960.

[4] M. Seabright, "A rapid banding technique for human chromosomes," *Lancet* ii, pp. 971-2, 1971.

[5] A. T. Sumner, H. J. Evans, and R. A. Buckland, "New Techniques for Distinguishing Between Human Chromosomes," *Nature New Biology*, vol. 232, no. 27, pp. 31-32, 1971.

[6] P.C. Nowell and D. A. Hungerford, "A minute chromosome in human chronic granulocytic leukemia," *Science*, 132, pp. 1197, 1960.

[7] ISCN (1995): *An International System for Human Cytogenetic Nomenclature*, Mitelman, F (ed); S. Karger, Basel, 1995.

[8] J. W. Gray and D. Pinkel, “Molecular cytogenetics in human cancer diagnosis,” *Cancer*, vol. 69, pp. 1536-1542, 1992.

[9] R. S. Ledley, F. H. Ruddle, J. B. Wilson, M. Belson, and J. Albarran. “The Case of Touching and Overlapping Chromosomes.” In: G. C. Cheng, R. S. Ledley, D. K. Pollock, A. Rosenfeld (eds), *Pictorial Pattern Recognition*, Thompson, Washington DC, 1968, pp. 87-97.

[10] K. R. Castleman. "Match Recognition in Chromosome Band Structure," *Biomed. Sci. Instrum.*, vol. 4, pp. 256-64, 1968.

[11] S. Kahan, T. Pavlidis, and H. S. Baird, “On the recognition of printed characters of any font and size,” *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 9, no. 2, pp. 274-288, 1987.

[12] B. Lerner, H. Guterman and I. Dinstein, “A Classification-Driven Partially Occluded Object Segmentation (CPOOS) Method with Application to Chromosome Analysis,” *IEEE Trans. on Signal Processing*, vol. 46, no. 10, pp. 2841-2847, 1998.

[13] G. Martin, “Centered-Object Integrated Segmentation and Recognition of Overlapping Handprinted Characters,” *Neural Computation*, vol. 5, no. 3, pp. 419-429, 1993.

[14] Y. W. Teh, *Learning to Parse Images*, University of Toronto, Toronto, Canada, Masters Thesis, 2000.

[15] S. MacLane and G. Birkoff, *Algebra*. Macmillan, 1979.

[16] G. Agam and I. Dinstein, "Geometric Separation of Partially Overlapping Nonrigid Objects Applied to Automatic Chromosome Classification," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 19, no. 11, pp. 1212-1222, 1997.

[17] J. Graham, "Resolution of Composites in Interactive Karyotyping," chapter in *Automation of Cytogenetics*, pp. 191-203, Springer-Verlag, Berlin, 1989.

[18] J. Serra, *Image Analysis and Mathematical Morphology*, Academic Press, London, 1982.

[19] D. Rutovitz, "Expanding Picture Components to Natural Density Boundaries by Propagation Methods. The Notions of Fall Set and Fall Distance," *Proc. Int. Joint Conf. Pattern Recognition*, 1978, pp. 657-664, Kyoto, Japan.

[20] J. M. Chassery and C. Gaybay, "An Iterative Segmentation Method Based on a Contextual Color and Shape Criterion," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 6, pp. 794-800, 1984.

[21] C. Garbay, "Image Structure Representation and Processing: A Discussion of Some Segmentation Methods in Cytology," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 8, pp. 140-146, 1986.

[22] L. Vanderheydt, F. Dom, A. Oosterlinck, and H. Van Den Berghe, "Two-Dimensional Shape Decomposition Using Fuzzy Subset Theory Applied to Automated Chromosome Analysis," *Pattern Recognition*, vol. 13, no. 2, pp. 147-157, 1981.

[23] Q. Wu, *Automated Identification of Human Chromosomes as an Exercise in Building Intelligent Image Recognition Systems*, Catholic University of Leuven, Leuven, Belgium, Doctoral Dissertation, 1987.

[24] L. Ji, "Intelligent Splitting in the Chromosome Domain," *Pattern Recognition*, vol. 22, no. 5, pp. 519-532, 1989.

[25] C. Arcelli and G. S. Di Baja, "A Width-independent Fast Thinning Algorithm", *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 7, no. 4, pp. 455-464, 1985.

[26] B. Lerner, H. Guterman, I. Dinstein, and Y. Romem, "Medial axis transform based features and a neural network for human chromosome classification," *Pattern Recognition*, vol. 28, no. 11, pp. 1673-1683, 1995.

[27] A. Moller, H. Nilsson, T. Caspersson, and G. Lomakka, "Identification of Human Chromosome Regions by Aid of Computerized Pattern Analysis," *Exp. Cell. Res*., vol. 70, no. 2, pp. 475-478, 1970.

[28] E. Granum, T. Gerdes, and C. Lundsteen, "Simple Weighted Density Distributions, WDDs for Discrimination between G-Banded Chromosomes," *Proc. Eur. Chrom. Anal. Workshop*, Edinburgh, Scotland, 1981.

[29] F. C. A. Groen, T. K. ten Kate, A. W. M. Smeulders, and I. T. Young, "Human Chromosome classification based on local band descriptors," *Pattern Recognition Letters*, vol. 9, no. 3, pp. 211-222, 1989.

[30] J. Gregor and E. Granum, "Finding Chromosome Centromeres Using Band Pattern Information," *Comput. Biol. Med*., vol. 21, no. 1-2, pp. 55-67, 1991.

[31] W. P. Sweeney, M. T. Musavi, and J. N. Guigi, "Classification of Chromosomes Using a Probabilistic Neural Network," *Cytometry*, vol. 16, pp. 17-24, 1994.

[32] P. A. Errington and J. Graham, "Application of Artificial Neural Networks to Chromosome Classification," *Cytometry*, vol. 14, pp. 627-639, 1993.

[33] B. Lerner, "Toward a Completely Automatic Neural-Network-Based Human Chromosome Analysis," *IEEE Trans. Trans. Systems, Man, and Cybernetics, Part B*, vol. 28, no. 4, pp. 544-552, 1998.

[34] S. O. Zimmerman, D. A. Johnston, F. E. Arrighi, M. E. Rupp, "Automated Homologue Matching of Human G-Banded Chromosomes," *Comput. Biol. Med*., vol. 16, no. 3, pp. 223-233, 1986.

[35] R. J. Stanley, J. M. Keller, P. Gader, and C. W. Caldwell, "Data-Driven Homogue Matching for Chromosome Identification," *IEEE Trans. Medical Imaging*, vol. 17, no. 3, pp. 451-462, 1998.

[36] L. Vanderheydt, A. Oosterlinck, J. Van Daele, and H. Van Den Berghe, "Design of Graph-Representation and a Fuzzy-Classifier for Human Chromosomes," *Pattern Recognition*, vol. 12, pp. 201-210, 1980.

[37] A. Carothers and J. Piper, "Computer-Aided Classification of Human Chromosomes: A Review," *Statistics and Computing*, vol. 4, no. 3, pp. 161-171, 1994.

[38] C. Lundsteen and J. Piper, *Automation of Cytogenetics*, Berlin, Springer-Verlag, 1989.

[39] D. Pinkel, T. Straume, and J. W. Gray, "Cytogenetic analysis using quantitative, high-sensitivity, fluorescence hybridization," *Proc. National Academy of Sciences of the United States of America*, vol. 83, pp. 2934-2938, 1986.

[40] P. M. Nederlof, S. van der Flier, J. Wiegant, A. K. Raap, H. J. Tanke, J. S. Ploem, and M. van der Ploeg, "Multiple fluorescence in situ hybridization," *Cytometry*, vol. 11, pp. 126-131, 1990.

[41] P. M. Nederlof, S. van der Flier, J. Vrolijk, H. J. Tanke, and A. K. Raap, "Fluorescence Ratio Measurements of Double-labeled Probes for Multiple in Situ Hybridization by Digital Imaging Microscopy," *Cytometry*, vol. 13, pp. 839-845, 1992.

[42] M. M. Le Beau, "One FISH, two FISH, red FISH, blue FISH," *Nature Genetics*, vol. 12, pp. 341-344, 1996.

[43] T. Ried, A. Baldini, T. C. Rand, and D. C. Ward, "Simultaneous visualization of seven different DNA probes by in situ hybridization using combinatorial fluorescence and digital imaging microscopy," *Proc. National*

Academy of Sciences of the United States of America, vol. 89, pp. 1388-1392, 1992.

[44] M. P. Sampat, K. Castleman and A. C. Bovik, "Pixel-by-Pixel Classification of M-FISH Images", *2nd Joint Conference of the IEEE Engineering in Medicine and Biology Society and the Biomedical Engineering Society*, October 23-26, 2002, Houston, TX, accepted for publication.

[45] T. Veldman, C. Vignon, E. Schröck, J. D. Rowley, and T. Ried, "Hidden chromosome abnormalities in haematological malignancies detected by multicolour spectral karyotyping," *Nature Genetics*, vol. 15, pp. 406-410, 1997.

[46] W. Schwartzkopf, B. L. Evans, and A. C. Bovik, "Minimum Entropy Segmentation Applied to Multi-Spectral Chromosome Images", *Proc. IEEE Int. Conf. on Image Processing*, Oct. 7-10, 2001, vol. II, pp. 865-868, Thessaloniki, Greece.

[47] W. Schwartzkopf, B. L. Evans, and A. C. Bovik, "Entropy Estimation for Segmentation of Multi-Spectral Chromosome Images", *IEEE Southwest Symposium on Image Analysis and Interpretation*, April 7-9, 2002, pp. 234-238, Santa Fe, NM.

[48] H. Stark and J. Woods, *Probability, Random Processes, and Estimation Theory for Engineers*, Upper Saddle River, N.J., Prentice Hall, 1994.

[49] P. M. Mather, *Computer Processing of Remotely-Sensed Images*, New York, John Wiley, 1987.

[50] http://www.adires.com/05/Project/MFISH_DB/MFISH_DB.shtml

[51] W. Schwartzkopf and K. Castleman, ``M-FISH Image Database'', *Cytometry*, in preparation.

[52] http://www.dssimage.com/cytoVision.html

Vita

Wade Carl Schwartzkopf was born March 2, 1975, to David and Karen Schwartzkopf in Indianapolis, IN. In 1997 he graduated *magna cum lauda* from Rose-Hulman Institute of Technology, receiving his B.S. with a double major in Computer Science and Electrical Engineering. While at Rose-Hulman, he held a National Merit Scholarship and a presidential scholarship for four years. In 1998 he received his M.S. in Electrical and Computer Engineering from The University of Texas at Austin. While at The University of Texas, he held several fellowships including the Microelectronics and Computer Development Fellowship (9/97-8/99), the College of Engineering Graduate Fellowship (9/97-5/01), the George J. Heuer, Jr. Ph.D. Endowed Graduate Fellowship (9/00-5/01), and the Charles W. Tolbert Endowed Presidential Scholarship (9/01-5/02). He has interned at Advanced Digital Imaging Research (6/00-8/00) where he worked on chromosome image analysis, and Hewlett-Packard Laboratories (6/01-8/01) where he researched automatic document layout. He has published work in the areas of pancreatic islet image segmentation, optical coherence tomography, two-dimensional phase unwrapping, neural networks, and chromosome image analysis. In addition, he has experience in object-oriented programming, signal processing, artificial intelligence, neural networks, and pattern recognition.

Permanent address: 3980 Diamond Lane, Indianapolis, IN, 46254

This dissertation was typed by the author.